

IJCAI2011

Workshop Proceedings of

The 6th International Workshop on Chance Discovery (IWCD6)

16 July, 2011

Room 1.4 of the New building,
Faculty of Mathematics and Philology, The Universitat de Barcelona,
Gran Via de les Corts Catalanes 585, Barcelona, Catalonia (Spain)

Organizers

Akinori Abe

NTT Communication Science Laboratories

3-1 Wakamiya, Morinosato, Atsugi-shi, Kanagawa 243-0198 JAPAN

E-mail: ave@ultimaVI.arc.net.my

Yukio Ohsawa

The University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 JAPAN

E-mail: ohsawa@sys.t.u-tokyo.ac.jp

Program Committee

David Bergner (NASA Ames Research Center, USA)

Mayumi Itakura Kamata (IBM Research-Tokyo, Japan)

Lorenzo Magnani (University of Pavia, Italy)

Peter McBurney (King's College, UK)

Ruediger Oehlmann (Kingston University, UK)

Wataru Sunayama (Hiroshima City University, Japan)

Xijin Tang (Chinese Academy of Science, China)

Shusaku Tsumoto (Shimane University, Japan)

Katsuyoshi Yada (Kansai University, Japan)

Contents

(1)	Preface.....	1
	Akinori Abe and Yukio Ohsawa	
(2)	Curation and Communication in Chance Discovery.....	3
	Akinori Abe	
(3)	Turning Down a Chance: An Argument From Simplicity.....	9
	Emanuele Bardone and Lorenzo Magnani	
(4)	Web-based Innovation Supporting System for Creative Ideas Emerging.....	15
	Wang Hao, Yukio Ohsawa, and Yoko Nishihara	
(5)	Discovering Probabilistic Models from ADS-B Data Sets using Digital Pheromones.....	21
	Kirk Ogaard and Ronald Marsh	
(6)	Discovering and Matching Service Providers among Heterogeneous Agents.....	27
	Hao Lan Zhang, Clement H.C. Leung, and Xijin Tang	
(7)	Extraction of Useful Utterances from Discussion Text for Improving Value of Idea.....	33
	Yoko Nishihara and Yukio Ohsawa	
(8)	Enhancing the Identification of Anomalous Events in Medicare Consumer Data Through Classifier Combination.....	39
	B.Sumudu.U. Mendis, D.Wayne Murray, Alison Sutinen, MingJian Tang, and Yingsong Hu	
(9)	Temporal Data Mining in Hospital Information Systems.....	45
	Shusaku Tsumoto and Shoji Hirano	
(10)	Discovering Health Problems in the Elderly using Data Mining Approaches learning.....	51
	Bogdan Pogorelc and Matjaž Gams	
(11)	Visualizing Systemic Risks and Predicting Systemic Financial Crises using Self-Organizing Maps.....	57
	Peter Sarlin and Tuomas Peltonen	
(12)	Map Interface for a Text Data Set by Recursive Clustering.....	63
	Wataru Sunayama, Shuhei Hamaoka and Kiyoshi Okuda	

Preface: The 6th International Workshop on Chance Discovery (IWCD6)

Akinori Abe (NTT Communication Science Laboratories)
Yukio Ohsawa (The University of Tokyo)

Since 2000 we have been organizing workshops on chance discovery in several conferences such as Int'l Conf. on Machine Learning (2005), Int'l Conf. on Data Mining (2010), and several invited sessions on chance discovery mainly in Int'l Conf. on Knowledge-Based and Intelligent Information & Engineering Systems. We also have had several special issues on chance discovery in several journals such as New Generation Computing (2003) and Int'l Journal of Advanced Intelligence Paradigms (2010).

For IJCAI, this is the first time to organize a workshop on chance discovery. Actually, we have organized workshops in ML and DM conferences. However, chance discovery is rather different from ML and DM. For general readers, we briefly introduce chance discovery. Chance discovery is the discovery of chance, rather than discovery by chance. A “chance” here means a novel or rare event/situation that can be conceived either as an opportunity or as a risk in the future. The “discovery” of chances is crucially important, since it may have a significant impact on human decision making. Desirable effects of opportunities should be actively promoted, whereas preventive measures should be taken in the case of discovered risks. In other words, chance discovery aims to provide means for inventing or surviving the future, rather than simply predicting the future. To explain “chance,” we usually use an example of economic “bubble” in Japan. A chance can be considered as an alarm, such as an inflation of money supply that will change the middle or long term economic “bubble” situation (Japan, in 1990). We can also find a similar symptom before the event, such as the rise in subprime loans that incited market panic in August 2007. Thus researches on chance discovery have been focusing on theories, strategies, methods and techniques of “how to discover chance?”

In this workshop, researches focusing on theories, strategies, methods and techniques of “how to discover chance?” are presented. A list of presenting papers are as follows:

- Akinori Abe: Curation and Communication in Chance Discovery.
- Emanuele Bardone and Lorenzo Magnani: Turning Down a Chance: An Argument From Simplicity.
- Wang Hao, Yukio Ohsawa, and Yoko Nishihara: Web-based Innovation Supporting System for Creative Ideas Emerging.
- Kirk Ogaard and Ronald Marsh: Discovering Probabilistic Models from ADS-B Data Sets using Digital Pheromones.
- Hao Lan Zhang, Clement H.C. Leung, and Xijin Tang: Discovering and Matching Service Providers among Heterogeneous Agents.

- Yoko Nishihara and Yukio Ohsawa: Extraction of Useful Utterances from Discussion Text for Improving Value of Idea.
- B.Sumudu.U. Mendis, D.Wayne Murray, Alison Sutinen, MingJian Tang, and Yingsong Hu: Enhancing the Identification of Anomalous Events in Medicare Consumer Data Through Classifier Combination.
- Shusaku Tsumoto and Shoji Hirano: Temporal Data Mining in Hospital Information Systems.
- Bogdan Pogorelc and Matjaž Gams: Discovering Health Problems in the Elderly using Data Mining Approaches.
- Peter Sarlin and Tuomas Peltonen: Visualizing Systemic Risks and Predicting Systemic Financial Crises using Self-Organizing Maps.
- Wataru Sunayama, Shuhei Hamaoka and Kiyoshi Okuda: Map Interface for a Text Data Set by Recursive Clustering.

Abe, and Bardone and Magnani give cognitive and philosophical aspect of chance discovery. They mainly discuss better chance selection and exhibition strategies from the cognitive and philosophical aspect. Zhang et al. discuss chance discovery from the viewpoint of an agent rank algorithm. This research discusses to select better chances. Ogaard and Marsh discuss a probabilistic model generation of pilot behaviour in aircraft operations. They discuss a strategy for less collision risk pilot. In fact, it is based on data mining techniques. However, it is a risk management system, which can be discussed in the context of chance discovery. Three papers (Hao, Ohsawa, and Nihihara, Sarlin and Peltonen, and Sunayama, Hamaoka, and Okuda) discuss the better visual interface for a decision making support. Actually, an interactive visualization has been one of the main theme in chance discovery. In addition, visualization can be discuss in the context of curation which has been one of the main theme in recent chance discovery workshops and sessions. Nishihara and Ohsawa discuss an influence of key utterance in discussions. They show such utterance is a chance to lead a discussion to the better direction. Three papers (Sumudu et al., Tsumoto and Hirano, and Pogorelc, Bosnic, and Gams) discuss problems in medical places and care houses. In such places, many problems which can be dealt with chance discovery will exist.

Thus this workshop will give current issues and future issues of chance discovery. In addition, readers of this proceedings will discover further research issues of chance discovery. We hope you will be joining the chance discovery research community.

Curation and Communication in Chance Discovery

Akinori Abe

NTT Communication Science Laboratories

Kanagawa, Japan

abe.akinori@lab.ntt.co.jp, ave@ultimaVI.arc.net.my

Abstract

In a previous paper, I proposed a concept of curation in chance discovery. Curation was discussed as a method of displaying chances for chance discovery. This paper views curation as communication, and chances can be discovered as a result of that communication. This type of communication can be regarded as affordance selection. In fact, for affordance selection, we communicate with an environment to select a proper affordance for an improved life or decisions. This type of communication is discussed as chance discovery where rare or potential events should be discovered to improve decision making. This type of communication is discussed as chance discovery where rare or potential events should be discovered for the better decision making.

1 Introduction

For more than 10 years, chance discovery has emphasised the importance of rare or novel events especially those, which would cause serious risk or generate benefits in the future; although data mining has recently focussed on abnormal events as outliers. Actually, the definition of a chance is: *“a novel or rare event/situation that can be conceived either as an opportunity or a risk in the future”* [18]. One of the key aims in chance discovery is to discover rare or novel events by (human-computer) interaction. With a proper interaction, rare or novel and important factors can be discovered, which may be hidden under frequently occurring events, noise, or complex relationships. In the context of chance discovery, various systems for suggesting the discovery of rare or novel events were proposed [4; 19]. In chance discovery, interactions between human and computer to aid the discovery process is one of significant themes. Such interactions could be achieved by providing a user-friendly interface. For instance, visualization systems for making users aware of unconscious preferences [9; 14], an analogy game which varies a construction of concepts according to perceptions, categorizations, and areas of focus derived from the expertise of the observer [16], a deposit overflow determination system to prevent various financial crises [24], ISOR-2, a combination of case-based reasoning

and statistical modeling system, which can deal with medical exceptions [23], and a web-based interactive interface which can check hidden or rare but very important relationships in medical diagnostic data sets [5] have been proposed in [4]. Those applications are real world applications where a discovery of chances plays an important role.

However, strategies of how to display chances have not been explicitly discussed in many applications. For a chance display strategy, I proposed a concept of curation for chance discovery [7].

A curation is a type of job in (art) museums and galleries, where curators remain up-to-date in the scholarly developments within their field(s), conduct original research and develop new scholarship that contributes to the advancement of the body of knowledge within their field(s) and within the museum profession as a whole [8]. In general, curators do not exercise communication with audiences except in special events such as gallery talks. Several applications which can be discussed from the viewpoint of curation were discussed in [7]. Those applications have an interactive interface or consider an interactive environment where users can access hidden information. Reviewing such applications, curation can be regarded as an interactive communication between curators and audiences. Thus, it will be necessary to discuss “curation” from the aspect of “communication.”

This paper discusses such communication as chance discovery and introduce

2 Curation in chance discovery

2.1 (Artistic) curation

In a report by the American Association of Museums Curators Committees (AAMCC) [8], it was pointed out that “curators are highly knowledgeable, experienced, or educated in a discipline relevant to the museum’s purpose or mission. Curatorial roles and responsibilities vary widely within the museum community and within the museum itself, and may also be fulfilled by staff members with other titles.” Then a definition of curation was given as follows;

- Remain current in the scholarly developments within their field(s); conduct original research and develop new scholarship that contributes to the advancement of the body of knowledge within their field(s) and within the museum profession as a whole.

- Make recommendations for acquiring and deaccessioning objects in the museum collection.
- Assume responsibility for the overall care and development of the collection, which may include artifacts, fine art, specimens, historic structures, and intellectual property.
- Advocate for and participate in the formulation of institutional policies and procedures for the care of the collection that are based on accepted professional standards and best practices as defined by AAM, CurCom, and other relevant professional organizations.
- Perform research to identify materials in the collection and to document their history.
- Interpret the objects belonging or loaned to the museum.
- Develop and organize exhibitions.
- Contribute to programs and educational materials.
- Advocate and provide for public use of the collection.
- Develop or contribute to monographs, essays, research papers, and other products of original thought.
- Represent their institution in the media, at public gatherings, and at professional conferences and seminars.
- Remain current on all state, national, and international laws as they pertain to objects in the museum collection.

In addition, AAMCC showed curatorial responsibilities as follows;

- A. Research, Scholarship, and Integrity
- B. Interpretation
- C. Acquisition, Care, and Disposal
- D. Collection Access and Use
- E. Replication of Objects in the Collection

Thus curators have responsibilities for various aspects of exhibition activities. However, the most important activity will be a plan of exhibition. For that the above activities such as research, interpretation and acquisition are necessary. They should properly exhibit a truth which is the result of their research and interpretations.

This can be called an old-style curation. In fact, as exercised in the exhibition “Bacon and Caravaggio” held in Museo e Galleria Borghese, Roma, Italy during October 2 2009 and January 24 2010, some curators do not insist their educational view.

For instance, Coliva wrote “This exhibition proposes a juxtaposition of Bacon and Caravaggio. It intends to offer visitors an opportunity for an aesthetic experience rather than an educational one... [10].” Then Coliva continued

“An exhibition of generally conceived and prepared with a historicist mentality, but when it materializes, the simultaneous presence of the works — in the sense precisely of their hanging — opens up parallels and poses very complex and spontaneous questions, which may even be unexpected and not all stem exactly from questions initially posed by art-historical motives and theses. There are parallels that appear by themselves to the visitor’s sensibility and are not imposed by a theory of the curator. This is certainly one aspect of the vitality

of exhibitions, which make the works live and in this are necessary for the works. The display itself, in the sense of the presentation of the works that appear in an exhibition — the spectacle of their being on display — creates trains of thought that are independent of the interpretations provided by art-historical scholarship. And since for a profound experience of understanding a work these ramifications sometimes are more surprising and significant than the achievements of a specialized scholarship in its own field of action, an art raised to the status of an enigma like Bacon’s seems to require the gamble of provoking these parallels. And since at the time, and again because of its qualitative greatness, Caravaggio’s art deserves a similar provocation, the juxtaposition thus satisfies a legitimate aesthetic desire. On the other hand, the juxtaposition is a modest and prudent solution, not so much for demonstrating, but for offering the attribute of “genius” — which the expressive common language attributes to the great artist of the past — opportunities to manifest itself. And the juxtaposition is induced by the Galleria Borghese itself, one of the most sensitive spaces with the simultaneous presence of genius.”

Although curators have a certain philosophy, they do not insist on their philosophy. Audiences will rather be able to discover additional meanings as well as the curators’ intended philosophy. This type of curation has the similar concept as chance discovery.

2.2 Does chance discovery need curation?

Definition: Chance discovery

Though in various articles, the definition of a “chance” is described which was introduced by Ohsawa [18], I wish to introduce it here again. In fact, it rather differs from the original definition in [18] to reflect the recent research interests.

A chance is a rare, hidden, potential or novel event or situation that can be conceived either as a future opportunity or a risk.

Then “chance discovery” research is a type of research to establish methods, strategies, theories, and even activities to discover a chance. In addition, it aims at discovering human factors for chance discoveries.

According to the above characteristic of chance discovery, I defined a new definition of curation in chance discovery as follows [7]:

- Curation is a task to offer users opportunities to discover chances.
- Curation should be conducted with considering implicit and potential possibilities.
- Chances should not be explicitly displayed to users.
- However, such chances should be rather easily discovered and arranged according to the user’s interests and situations.

- There should be a certain freedom for user to arrange chances.

Thus a part of the curatorial tasks functions as a part of a mechanism of chance discovery. That is, non-educational curation sometimes takes the role of assisting chance discovery.

3 Curation and communication

In the previous section I pointed out the importance of non-educational curation for assisting chance discovery. I introduced some successful curations in [7]. However, such successful curations are rather rare. Usually, curations are based on an educational system and it is rather difficult to discover rare issues. Such a system provides fixed information and therefore by following such information, we can reach the correct and intended goal. If we do not follow the information, we cannot reach the correct goal.

Advertising communication also aims at offering certain information to audiences. Sometimes such information is implicitly expressed in an advertisement. As Pop summarized [21], advertising communication relies considerably on inferences and assumptions which help in proceeding towards eventual interpretations. Based on Grice's seminal theory of cooperative communication (cooperative principles, CP) and inferencing through a maxim of "filling in" or/and flouting, different interpretations could be accommodated by the linguistic theory. Pop extended her discussion by introducing Relevance Theory (RT)[22], which explains hidden and additional information in advertisement.

Both curation and advertisement intend to deliver a certain concept to audiences. As easily guessed "communication" plays a significant role in curation and advertisement.

In the following I will review chance discovery from the aspect of communication. Actually, the following application is affordance, which can be regarded as communication between human and environment. We communicate with an environment and ecologically selected affordance which will be better or best for our lives.

3.1 Information offering strategies for dementia persons

In [3], based on abduction, I formalized the concept of an affordance-based support system for dementia patients. For a proper and extended usage of an object, it is necessary to present appropriate information about it. However, for dementia patients, normal information offering strategies cannot function. Therefore, it is necessary to prepare appropriate information offering strategies especially for dementia patients. However, for a progressive and promising system, it is not realistic to prepare all the necessary information about items. Sometimes such information is not correct and will change in the future. For instance, it is ridiculous to attach a sign such as "You can sit here." to tree stumps. It is rather realistic to suggest information about its hidden functions. Such hidden information can be presented as certain stimuli in such situations, because, even dementia patients sometimes perform better, if they receive certain stimuli. The question

is what type of stimulus should be presented and how can the recognition of a stimulus be achieved.

For that I introduced the concept of affordance which was ecologically introduced for perceptual phenomena by Gibson [11; 12]. Gibson defined the affordance of something as "a specific combination of the properties of its substance and its surfaces taken with reference to an animal."

Actually such stimulus should be "afforded" by the user. That is, it can be regarded as an "affordance" in an environment. Accordingly we introduce the concept of affordance to a dementia care system. Proper affordance might give some support to a dementia patient's understanding (meaning) of objects. When meaning is fixed, by using the abduction framework, the situation of determining affordance will be logically described as follows:

$$Object \cup affordance \models meaning \quad (1)$$

$$Object \cup affordance \not\models \square \quad (2)$$

That is, affordance can be regarded as a hypothesis. We can select a consistent affordance (equation (2)) in the environment (hypothesis base) to explain meaning. In addition, for understanding a subset of similar afforded objects ($Object'$), the affordance determination situation will be logically described as follows:

That is, affordance can be regarded as a hypothesis. We can select consistent affordance (equation (2)) in the environment (hypothesis base) to explain meaning. In addition, for understanding a subset of similar afforded objects ($Object'$), the affordance determination situation will be logically described as follows:

$$Object \cup Object' \cup M \cup affordance \models meaning \quad (3)$$

M is a mapping function [13] from $Object$ to $Object'$. That is, to understand the same meaning of the subset of or similar afforded objects, an additional mapping function M is required. Thus if M can be determined and the usage of $Object$ is known, $Object'$ can also be understood. In fact, for healthy individuals, M is easy to understand. However, for dementia patients, it is rather difficult to understand and to determine M . Affordance can be formalized by an abduction framework.

3.2 Advertising communication and curation

I will not discuss details of advertising communication in this paper due to its length limitations. Some suggestions were described in [1]. These were not discussed from the perspective of affordance or abduction, but of advertising communication. Frequently an advertisement planner tries to hide an actual intention in phrases or pictures used in advertisements. This type of strategy can be discussed from the aspect of communication. In addition, it can be said that advertisements offer a type of affordance to be selected by audiences. Of course, such affordance will be effectively selected by a suitable curatorial tasks.

4 Communication, abduction and affordance

In the previous section, I briefly introduced my previous work based on affordance which can also be discussed as an application of curation. For the application, the most important issue is how to suggest hidden information as affordance. My assumption was that complex situation can be transformed to a combination of simple situations. This type of combination can be achieved as curation. For such system, curatorial strategy should be introduced to offer understandable mapping

I formalized affordance selection with abduction and analogical mapping. In addition, analogical mapping can also be executed in an abduction framework. An affordance selection is achieved by communication between environment and humans. If it is difficult, it means that communication between environment and humans is missing. For an appropriate affordance selection, it is necessary to perform suitable communication between an environment and an individual. In this section, I will view missing communication as chance and propose a simple strategy to determine a chance.

4.1 Missing communication as chance

I have previously analyzed nurses' activity pattern by using *Kamishibai – KeyGraph*[®] [2].

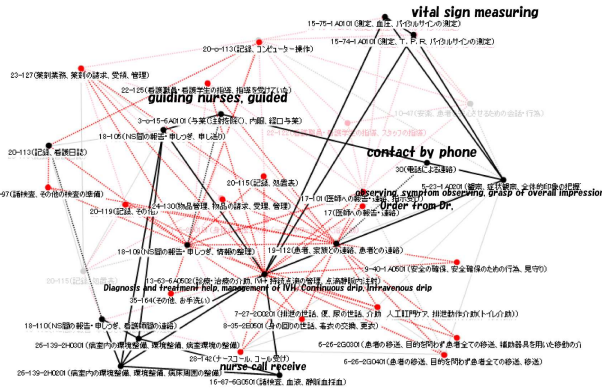


Figure 1: Result from *Kamishibai – KeyGraph*[®]

When we analyze nurses' activities by applying *Kamishibai – KeyGraph*[®], a result shown in Fig. 1 is obtained. For an expert of *KeyGraph*[®], it may not be so difficult to determine the point of chance. However, usually this type of task is conducted by both a *KeyGraph*[®] expert and an expert in the field (nurse, doctor or risk manager in this case). Of course, a good collaboration sometimes produces novel viewpoint which can also be regarded as a chance. However, it is not always the case that both experts can work together. The preferable situation is given by a field expert who can properly control such an analysis system to exercise chance discovery.

To go back to the application of affordance, the figure from *KeyGraph*[®] should offer affordances, which the user will be aware of and will select. That is, in the case of Fig. 1, affordances come from several key points which are candidates of chance points as well as the other points. For a suffi-

cient result, users will select proper affordances. For instance, for professional observer, Fig. 1 might be observed as Fig. 2 where light yellow regions contain candidate affordances. For an untrained individual, such observation will be rather difficult. Therefore suggested markings as those in Fig. 2 will be helpful. However, of course they should be correct and as pointed out before they should not indicate intentions, which may bias the observer. They should rather be displayed naturally and should not disturb or control the user's thought process.

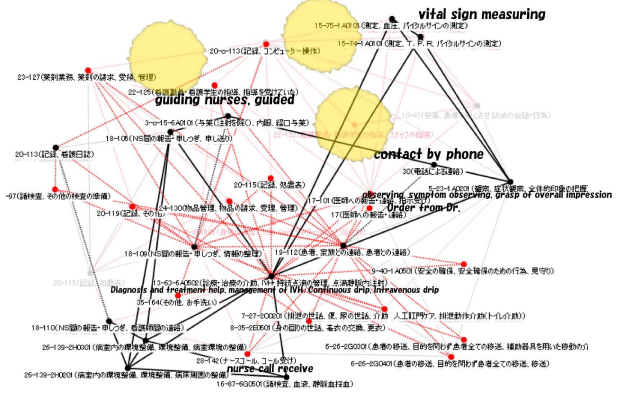


Figure 2: Result from *Kamishibai – KeyGraph*[®] with suggestions

4.2 How to determine missing communication?

In [6], we proposed an=A0interactive support system to suggest chance points, which can analyze nurses' conversation to determine points at which communication might be missed. It is tuned to communication analysis. However, a fundamental concept of the system is based on abduction. That is, the system back-traces communication paths from the point where an accident or incident is noticed.

In general, abduction functions to determine reasons (explanations) for an observation or sometimes for missing events to explain an ideal observation. In fact, as shown above, in the information offering system for dementia patients [3], I applied abduction for generating a mapping for understanding affordance. Thus by adopting the abduction framework, it will be able to suggest missing communication, which can be regarded as a set of affordances. If we adopt a graphical system, for instance *KeyGraph*[®], it is necessary to build a set of logical statements from the result. After that abduction can be applied to abduce missing communication, which will be an affordance. Of course, many candidate affordances will be abduced. However, logically certificated affordances will be helpful for the further decision making.

As Magnani and Bardone discussed, the problem of an activity promoting fake chances¹ will usually occur during the activity of chance-seeking [15]. A bullshit has no intention to cheat and to be logically true. Accordingly, it is rather

¹They call it bullshit.

difficult to determine it as a fake, because the above procedure seems to be guaranteed by aq logical procedure. In order to ignore such “bullshit,” curation will play a important role. That is, by a appropriate curation such “bullshit” will be shown as lie (=“bullshit.”) For that, curation as communication is very important. During such communication, users will be guided into a proper direction.

In this section, I used an example output by *Kamishibai – KeyGraph*[®]. Abduction can be applied to various applications, so the above strategy can be applied to any application which has a causation. Thus, a proper strategy based on curation (abduction, affordance, and communication) should be developed for an improved chance discovery.

5 Conclusions

In this paper, I briefly reviewed curation in chance discovery which I have proposed in [7]. In fact, for chance selection, curatorial strategy is very important. For visualization, various applications have been proposed. However, for such applications as *Kamishibai – KeyGraph*[®], it is rather difficult to discover chances if the user is not an expert.

For a good interface, Norman claimed that “make it easy to determine what actions are possible at any moment (make use of constraints).” [17] In addition, Norman claimed that “make things visible, including the conceptual model of the system, the alternative actions and the results of actions.” Norman’s concept is based on a part affordance. However, Norman’s point is rather different from what I discussed in this paper. Norman’s main point is to design actual tools, for instance a door, to be obviously recognized by everybody as a thing which should be opened or closed by pulling or pushing a certain mechanism (doorknob). The mechanism should be easy to understand. For a good interface, it will be the best strategy. However, as a chance discovery’s interface, it will steal a certain chance. For instance a door might be used as a BBQ grill board. A doorknob might be used as a hummer.

Of course Norman’s visible interface concept is correct in an actual or usual situation. However, for chance discovery the other “visible” interface is necessary. A strategy I introduced in [3] will be one of “invisible” interfaces under the concept of chance discovery. The other strategy I suggested in the previous section will be one of “semi-visible” interface under the concept of chance discovery. Where users can discover potential element (affordance) that is correct but is not obviously considered or controlled by a system designer. This type of suggestion is very important in chance discovery. Logically guaranteed candidates (affordance) can be displayed by abductive procedure. Users can determine which candidate is better for their (future) decision making. For such procedures, a strategy by abduction and analogical mapping will be effective. Abduction functions as a logical guarantee and analogical mapping functions as an extension of or support for potential element discovery. In addition, in this paper, I considered the introduction of a concept of communication in the above procedure. In affordance selection, communication between an environment and an individual will be conducted. I regarded a missing communication as a chance. In general, affordance is easy to obtain if one is in an en-

vironment and feels the environment. For chance discovery, affordance is not easy to obtain, because it potentially exists in an environment. There will be a missing link between the environment and us. In order to select the better affordance, it is necessary to discover such a missing link. A missing links can be suggested by abduction and extended or suited by an analogical mapping procedure which can also be achieved by abduction.

References

- [1] Abe A.: Language Sense, *Proc. of PRICAI2004 Workshop on Language Sense on Computer*, pp. 1–9 (2004)
- [2] Abe A., Ohsawa Y., Ozaku H.I., Sagara K., Kuwahara N., and Kogure K.: Communication error determination model for multi-layered or chained situations, *Proc. of PAKDD 2008 Working Notes of Workshops on Data Mining for Decision Making and Risk Management*, pp. 305–316 (2008)
- [3] Abe A.: Cognitive Chance Discovery, in *Universal Access in Human-Computer Interaction — Addressing Diversity* (Stephanidis eds.), *UAHCI 2009, Held as Part of HCI International 2009, Proceedings, Part I (LNCS5614)*, pp. 315–323, Springer-Verlag (2009)
- [4] Abe A. eds.: Special issue on Chance Discovery, *International Journal of Advanced Intelligence Paradigms*, Vol. 2, No. 2/3 (2010)
- [5] Abe A., Hagita N., Furutani M., Furutani Y., and Matsuoka R.: An Interface for Medical Diagnosis Support —from the viewpoint of Chance Discovery, in [4], pp. 283–302 (2010)
- [6] Abe A., Ohsawa Y., Ozaku I.H., Sagara K., Kuwahara N., Kogure K.: Communication Error Determination System for Multi-layered or Chained Situations, *Fundamenta Informaticae*, 98, pp. 123–142 (2010)
- [7] Abe A.: Curation in Chance Discovery, *Proc. ICDM2010 5th International Workshop on Chance Discovery*, pp. 793–799 (2010)
- [8] American Association of Museums Curators Committee: A code of ethics for curators, http://www.curcom.org/_pdf/code_ethics2009.pdf (2009)
- [9] Amitani S. and Edmonds E.: A Method for Visualising Possible Contexts, in [4], pp. 110–124 (2010)
- [10] Colvia A.: Caravaggio Beckons to Bacon: The Beauty of Sorrow, in *Caravaggio Bacon (A. Coliva and M. Peppiatt eds.)*, pp. 17–22 (2009)
- [11] Gibson J.J.: The Theory of Affordances, *Perceiving, Acting, and Knowing* (Shaw R. and Bransford J. eds.) (1977)
- [12] Gibson J.J.: *The Ecological Approach to Visual Perception*, Houghton Mifflin (1979)
- [13] Goebel R.: A sketch of analogy as reasoning with equality hypotheses, *Proc. of Int’l Workshop Analogical and Inductive Inference (LNAI-397)*, pp. 243–253 (1989)

- [14] Maeno Y. and Ohsawa Y.: Reflective visualization and verbalization of unconscious preference, in *[4]*, pp. 125–139 (2010)
- [15] Magnani L. and Bardone E: Faking Chance: Cognitive Niche Impoverishment, Proc. KES2010(LNAI6278), Part III, Springer Verlag (2010)
- [16] Nakamura J., Ohsawa Y., Nishio H: An analogy game: toward cognitive upheaval through reflection-in-action, in *[4]*, pp. 220–234 (2010)
- [17] Norman D.: *The Design of Everyday Things*, Addison Wesley (1988)
- [18] Osawa Y. and McBurney P. eds: *Chance Discovery*. Springer Verlag (2003)
- [19] Ohsawa Y. and Tsumoto S. eds: *Chance Discoveries in Real World Decision Making*, Data-based Interaction of Human Intelligence and Artificial Intelligence Series: Studies in Computational Intelligence, Springer Verlag (2006)
- [20] Ohsawa Y., Ito T., and Kamata M.: Kamishibai Key-Graph as Scenario Map Visualizer for Detecting Transient Causes from Sequential Data, *Proc. of PAKDD 2008 Working Notes of Workshops on Data Mining for Decision Making and Risk Management*, pp. 272–283 (2008)
- [21] Pop A.: COVERT COMMUNICATION IN ADVERTISING: A CASE STUDY, [http://www.upm.ro/facultati_departamente/stiinte_litere/conferinte/situl_integrare_europeana/Lucrari2/Anisoara Pop.pdf](http://www.upm.ro/facultati_departamente/stiinte_litere/conferinte/situl_integrare_europeana/Lucrari2/Anisoara%20Pop.pdf)
- [22] Sperber D. and Wilson D.: *Relevance (2nd. ed.)*, Blackwell (1995)
- [23] Vorobieva O. and Schmidt R.: Case-Based Reasoning to Explain Medical Model Exceptions, in *[4]*, pp. 271–282 (2010)
- [24] Yada K., Washio T., and Ukai Y.: Modeling Deposit Outflow in Financial Crises: Application to Branch Management and Customer Relationship Management, in *[4]*, pp. 254–270 (2010)

Turning Down a Chance: An Argument From Simplicity

Emanuele Bardone and Lorenzo Magnani

Department of Philosophy and Computational Philosophy Laboratory

University of Pavia, Italy

bardone@unipv.it and lmagnani@unipv.it

Abstract

The notion of bullshit has quite recently acquired a theoretical and philosophical dignity. Described as the careless attitude that an agent has towards the truth-value of what he believes in, bullshitting is a valuable candidate for shedding light on those situations in which a chance should be turned down inasmuch as it is a fake chance. By contrasting the idea of bullshitting with that of scientific rationality, we will contend that the former favors what we call *cognitive baroquism*. That is, bullshitting has the major effect of unnecessarily complicating things and so intoxicating our cognitive niches with fake chances. In order to capture this feature of bullshitting as chance-faking, we will develop an argument from simplicity. Coming from the tradition of philosophy of science, the notion of simplicity might be of great help in spelling out the main effects of chance-faking and, at the same time, it gives us a valuable heuristic to understand when a chance should be rather turned down than actively pursued and implemented.

1 Maintaining Cognitive Niches and the Problem of Detecting Fake Chances

As defined by Ohsawa and McBurney [Ohsawa and McBurney, 2003], a chance is a new event or situation conveying both an opportunity and a risk in the future. Recently, a number of contributions have acknowledged the abductive dimension of seeking chances with relation to science [Magnani, 2005; Magnani and Bardone, 2008; Magnani, 2009; Abe, 2009]. As maintained by Magnani and Bardone [Magnani and Bardone, 2008] and Abe [Abe, 2009], the process of chance detection (and creation) is resulting from an inferential process – mainly abductive – in which the agent exploits latent clues and signs signaling or informing the presence of an action opportunity [Magnani and Bardone, 2008; Bardone, 2011]. In this case, as argued by Magnani [Magnani, 2009] the abductive inferential dimension has to be considered beyond its sentential/computational one.

Accordingly, an inference is a form of sign activity in which the word sign encompasses several types of sign, for

instance, symbol, feeling, image, conception, and other representation [Peirce, 1931 1958, 5.283]. Moreover, the process of inferring – and so the activity of chance seeking and extracting – is carried out in a distributed and hybrid way [Magnani, 2009]. This approach considers cognitive systems in terms of their environmental situatedness: instead of being used to build a comprehensive inner model of its surroundings, the agent's perceptual capacities are seen as simply used to obtain “what-ever” specific pieces of information are necessary for its behavior in the world. The agent constantly “adjusts” its vantage point, updating and refining its procedures, in order to uncover a piece of information. This resorts to the need of specifying how to efficiently examine and explore and to the need of “interpreting” an object of a certain type. It is a process of attentive and controlled perceptual exploration through which the agent is able to collect the necessary information: a purposefully moving through what is being examined, actively picking up information rather than passively transducing [Thomas, 1999]. In this sense, humans like other creatures are ecological engineers, because they do not simply live their environment, but they actively shape and change it looking for suitable chances, epistemic for example, like in the case of scientific abductive thinking.

Generally speaking, the activity of chance-seeking as a plastic behavior is administered at the eco-cognitive level through the construction and maintenance of the so-called cognitive niches. The various cognitive niches humans live in are responsible for delivering those clues and signs informing about an (environmental) chance. So, the mediating activity of inferring as sign activity takes place (and is enhanced) for the presence of the so-called eco-cognitive inheritance system [Odling-Smee *et al.*, 2003]. That is, humans can benefit from the various eco-cognitive innovations as forms of environmental modifications brought about and preserved by the previous generations. Indeed, many chance-seeking capacities are not wired by evolution, but enter one's behavioral repertoire because they are secured not at the genetic level, but at the eco-cognitive one – in the cognitive niches. The second important point to mention is that humans as chance extractors act like eco-cognitive engineers [Magnani and Bardone, 2008; Bardone, 2011]. Accordingly, they take part in the process of extracting chances by performing smart manipulation in order to turn an external constraint into a part of their extended cognitive system.

The persistence of cognitive niches is granted by various activities of cognitive niche maintenance whose main role is to preserve the integrity of our cognitive repertoires. So, humans as eco-cognitive engineers are also chance-maintainers [Bardone, 2011]. Humans act so as to preserve those chances (and “chances of chances”) that have been proved as successful and worth pursuing. That is what allows us to have an eco-cognitive inheritance that, in turn, enriches or even creates the behavioral chances a community or a group of people has for surviving and prospering.

Innovation management can be considered an activity of cognitive niche maintenance, as it deals with how to enrich a pre-existing cognitive repertoire. Technological innovation is indeed part of the process in which our cognitive niches are progressively enriched with new chances at several levels, namely, motor, sensory, and cognitive [Bardone, 2011]. However counterintuitive it might sound, not all improvements due to technological innovation are really enhancing. Not all of them can really enrich our cognitive repertoire with new and valuable chances. Some technologies may not be enriching at all. Conversely, they may contribute to introducing chances that intoxicate or poison pre-existing cognitive niches. Such chances are *fake* chances that should be turned down. Now the question is: when should we turn down a (fake) chance, notwithstanding the fact that it is a product of technological innovation?

In investigating chance-faking, we posit that pointing at the very roots of scientific rationality is extremely fruitful. As a matter of fact, the unprecedented success that human beings have achieved in the last centuries as eco-cognitive engineers bears an essential connection to the progress of science and scientific rationality. The latter does not only regard with the results of science; it embraces all those habits – epistemic, but also moral, social, and political ones – which are prerequisites for the development and progress of science in society. Generally speaking, scientific rationality is characterized by all those methods and forms of reasoning that consider truth as the value one should not compromise on. Indeed, the kind of truth we are talking about has a specific meaning to science and within scientific practice.

We will start our discussion with a philosophical account of bullshitting provided by Harry Frankfurt. If the hallmark of scientific rationality is a strong commitment towards truth, bullshitting is conversely characterized by the absence of such commitment. The absence of commitment towards truth makes bullshitting the perfect spot from which we can look at the phenomenon chance-faking.

By contrasting the notion of bullshitting with that of scientific rationality, we will contend that bullshitting favors what we call *cognitive baroquism*. That means that bullshitting has the major effect of unnecessarily complicating things. In order to capture this feature of bullshitting, we will develop an argument from simplicity. More precisely, we will show that the very notion of simplicity – coming from the tradition of philosophy of science – might be of help in spelling out the main effects of chance-faking.

In the last part of the paper we will apply the idea of cognitive baroquism to technology. We will describe two cases in which bullshitting may affect the way we use technology

and, as a result, facilitate the proliferation of fake chances.

2 What Bullshitting Is and Its Philosophical Relevance

The notion of bullshit introduced by Frankfurt [Frankfurt, 2005] may help describe a fundamental feature of people: when they exhibit a carelessness about truth and can easily perpetrate violence – favoring deception and fraud – just thanks to this systematic undervaluing of truth [Misak, 2008], at least when we consider truth as bounded up with the concepts of reasons, evidence, experience, inquiry, reliability, and standard or rational belief, for example in agreement with a kind of scientific mentality.

According to Frankfurt, there is an important distinction to make between a bullshitter and a liar. The difference between the two is that the liar has a general concern about truth. And this is just because, in order to tell a lie, he has to know what the truth is. Although the liar fails to be cooperative with respect to a certain state of things in the world, he is indeed cooperative with respect to his attitude towards truth. The fabrication of a lie may require a great deal of knowledge and it is mindful: putting it simple, it requires the guidance of truth. More generally, a certain state of mind – namely, an intention to deceive – is required to the liar while making a statement. This attitude is what makes his statement potentially informative. For instance, consider the case of a person telling us that he has money in his pocket, when he has not. His lie is informative as one can guess whether he lied or not. What is interesting about lying is that there always is a reason why a person may not tell the truth: lies and deceit can be detected. People, for instance, have at disposal both verbal and non verbal cues enabling them to detect potentially deceiving situations [Vrij, 2008]. A minor detail about the way a person dresses may be a chance suggesting a man that his wife is cheating on him, and vice-versa. Sometimes people fear the consequences of knowing the truth, therefore they prefer not to investigate. But this does not mean that they would not succeed. Quite the opposite. So, given the fact that a liar is committed to the truth-value of one’s belief, lying may contribute to hiding a chance, but at the same time it might be revealing it, as one detects the deceiving intention of the liar.

According to Frankfurt, the case of bullshit is different, as the bullshitter is supposed to lack any concern or commitment to the truth-value of what he says. What turns out to be extremely puzzling is not the content, but his attitude. For instance, as we have already pointed out above, a liar voluntarily makes a wrong claim. But in doing so he conveys a certain commitment to the truth-value of what he claims. A bullshitter does not care about it. As just mentioned, a liar has a deceptive intention that can be detected. Whereas the case of bullshitter is different. When a person believes *P*, he intends to believe *P*. And that intention becomes meaningful to other people. In the case of bullshitter, he believes without a real intention to believe what he believes. So, what really defines a bullshitter is his attitude towards truth: he fails to be committed to truth. He simply does not care whether what he says is true or false, accurate or inaccurate.

3 Bullshitting as Chance-Faking

The philosophical illustration of bullshit we have presented so far allows us to argue that bullshitters are basically *chance-fakers*. Why are they so? The deceiving character of chance-faking is related to the fact that one acts as if a chance were present, when it is not. Roughly speaking, what comes out of the bullshitter's mouth is *hot air* or *vapor*, meaning that the informative content transmitted is *nil*.

The deceiving character of chance-faking is particularly evident in the case of those cognitive processes involving a collaborative dimension, for instance, the case of second-hand knowledge, that is, the possibility of passing a chance on to another person. But how does the chance-faker engage other people in discussion, for instance?

As argued by Simon, we do lean on what other people say. That disposition to rely on social channels in problem-solving activities is what he called *docility* [Simon, 1993]. People tremendously benefit from aids and resources provided by their fellows. This has a major cognitive advantage: that almost any body can trust other people and so have at disposal chances that, first of all, he has never personally experienced, and, secondly, that are already available to pick up. That is one of the most important assets describing cognitive economy, that is, the need to reach a sort of trade-off between accuracy of a decision and the limited time one is bounded to. Indeed, trust is not informatively empty. One decides to trust another person, because he has reasons to do so. There are a number of clues we make use of in order to consider a particular source of information (a person, for instance) as trustworthy or not. For instance, people usually tend to trust people that exhibit some authority [Jackson, 2008]. In our opinion, the chance-faker basically provides his partner in conversation with chances that are basically faked. To make this point clearer, we introduce the notion of *frame*.

Usually, when entering a discussion a person is supposed to have some expectations. For example, he wants to gain a favor or he wants to display a certain identity. On some other time, he wants to express his dominance over the other or he simply wants to entertain himself or others taking part in the conversation. Usually, all of these these reasons lead a person to *frame* what he is doing in a particular way rather than another depending on what his purpose is [Goffman, 1986; Hample and Young, 2009]. Now, the question is: what kind of *frame* – to use Goffman's terminology – does the chance-faker select? What does he hold in his mind? What purpose?

Our take is that the chance-faker speaks without planning and doing so he simply *blurts*. As Hample put it, blurring “comes simply out of cognition without alteration” [Hample and Young, 2009, p. 23]. Basically, he acts as if he responded *unreflectively* and *mindlessly* to an invitation to have his say – no matter whether real or imaginary. Instead of admitting that he has nothing to say (and hence holding his peace), the chance-faker enters a debate or discussion without questioning if he has actually something to say. It seems that he suffers from a sort of pathological certainty, meaning that he does not come to doubt when he should about his competences and knowledge. So, his pathological certainty may produce what could be called *pathological opining*: he cannot abstain from

opining about topics clearly beyond his knowledge.

We have a last remark on frames and bullshitting. As already mentioned, a person entering a discussion usually bears in mind a certain understanding of what is going on. That is, he (more or less tacitly) interprets why others are arguing, i.e. their frame. Indeed, it is important to understand what frame a partner is using during a conversation: it avoids misunderstandings and, upon occasion, facilitates cooperation and information exchange. By contrast, when two or more people do not share the same frame, cooperation is less likely to come about. One can even project a false frame for his partner(s) and then hack the conversation to have a private gain. When this happens, Hample *et al.* argue that the victim “is said to be *contained*” in the swindler's frame [Hample *et al.*, 2009]. We contend that something similar happens with the chance-faker. We argued that the chance-faker has no interest or purpose in his mind when entering a conversation, but he simply responds to a (sometimes imaginary) invitation like reacting to a stimulus (for instance, “rate this product!”). Hence, he actually does not project any frame but, insofar as he is engaged in a conversation, he gives the impression he is employing a frame. It is worth noting here that the deceptive nature of bullshitting does not rest on the projection of a false frame, but on the fact that his victims are naturally inclined to assume that he actually projects a frame when he does not. The chance-faker's conversational contribution is nil, but nevertheless he is very often taken seriously, say, as if he really contributed.

The major consequence is that an imaginary frame is projected and the chance-faker's victims get into it and thus are *contained*, just like in the case of a fraudulent attack. Indeed, once they realize that the frame was merely imaginary, they may experience all that as fraudulent and aggressive. As a consequence, the chance-faker may even try to justify his misbehavior, for instance, saying that he did not really mean to trick his conversation partners, but that he just wanted to play. This sort of post-projection does not solve the problem. A fake chance is fabricated and it is available to be picked up. In fact the newly-projected meaning sharpens the dissonance felt by his victims. According to Goffman, when a certain activity is re-framed, or *re-keyed*, it is said that a *lamination* (or layer) is added to that activity [Goffman, 1986, p. 82]. More precisely, a lamination is a new meaning a person adds up to pre-existing layers. Indeed, the prerequisite for a lamination to be perceived is that the partners in conversation apply the appropriate frame. So, a lamination does not conflict or erase all the pre-existing meanings that a certain activity – for instance, arguing – may have. But it somehow becomes a new layer that metaphorically *laminates* it providing new interpretations, but, at the same time, it also favors the emergence of dissonance, in case the participants are not able to successfully re-key the activity.

4 Chance-faking as Cognitive Baroquism and Its Application to Technology

As already discussed, bullshitting as chance-faking is a kind of attitude characterized by the absence of commitment towards the truth-value of what one believes in. More gener-

ally, chance-faking deals with all those situations in which a person (or a group of persons) displays a careless attitude concerning how to generate and/or select hypotheses, which eventually gives shape to his beliefs. In this sense, insofar as the chance-faker is involved in generating or selecting hypotheses, he embraces a degraded form of abductive cognition.

The careless attitude toward the whole process of hypothesis generation and selection – which is responsible for belief formation – may be manifest in various ways. Indeed, the most distinctive trait characterizing the poor abductive performance of a chance-faker is that he does not pay attention to evidence. But there is one aspect – to a large extent ignored – which drastically contrasts, for instance, with scientific rationality: bullshitting – especially in its more sophisticated versions – tends to what we call *cognitive baroqueism*. That is, chance-fakers do not have any commitment to simplicity, which, conversely, is peculiar to scientific rationality. We posit that chance-fakers do not take into serious consideration the truth-value of the assumptions they implicitly make. As mentioned above, the kind of truth we are talking about is specific to the scientific discourse and practice, in which a subset of values are actively sought out. We are now going to show how this attitude – the lack of commitment to simplicity – might have negative effect in chance-seeking activities.

Bullshitters fail to exhibit any commitment to the truth-value of what they say, but they successfully manage to say something. Bullshitters always have an answer, no matter what they are talking about, as they rely on *pathological opinionating* [Bardone, 2011]. By this term we mean that they does not come to doubt when they should about their competences and knowledge: they cannot abstain from opining about topics clearly beyond his knowledge. Their careless attitude towards any procedure of belief formation and revision allows them to say something about everything *anyway*. In this sense, we posit that chance-fakers *mimic* what philosophers of science call *consilience*.

Consilience refers to how much a theory explains – its explanatory breadth [Thagard, 1992]. Generally speaking, the more consilient a theory is, the more it explains. Indeed, the kind of consilience we are talking about is not the same of a scientific theory or hypothesis. Let us say that we rather refer to a *pragmatic* form of consilience than an epistemic one. That is, the ability in a discussion of formulating opinions which allow to jump from one topic to another preserving a certain degree of unity and generality. In this sense, the kind of consilience exhibited by the chance-faker is similar to the one of a *know-it-all*. However, we claim that chance-faker's pragmatic consilience is achieved at the expense of simplicity. We will now illustrate why.

The medieval philosopher William of Ockham pointed out that it is worthless to do with more what can be done with less. In philosophy of science simplicity plays a pivotal role in comparing two or more theories. As defined by Thagard, simplicity refers to the number of kinds of entities (hypotheses, concepts, etc.) that a given theory actually assumes in order to explain certain facts [Thagard, 1978]. The less entities a theory assumes, the simpler it is. More generally, simplicity points out that a theory should make use of as less entities

as possible. Now, we analogically argue that a chance-faker is completely careless about what he should assume in order his belief to be consistent and epistemically grounded. That is, he can easily resort to pathological opinionating insofar as he completely dismisses any commitment towards the truth-value of what his beliefs would imply to be acceptable – the auxiliary hypotheses in the case of a scientific theory, for instance. The consequence is the dismissal of simplicity.

What we are going to do now is to try to show how this particular cognitive distortion due to chance-faking may be fruitfully applied to technology. More precisely, we will show that the chance-faking attitude we have just described can also be applied to technology, and so give us interesting hints about turning down a chance.

As already pointed out, bullshitting is intrinsically related to the way humans form and select their beliefs. But in a way – and this is our major claim – the same attitude might be exhibited with relation to the construction, selection, and evaluation of technologies and the chances they provide. In order to make this clear, we maintain that simplicity and consilience may fruitfully be part of our toolbox for detecting and avoiding what we called fake chances. Indeed, the role of consilience and simplicity for comparing two or more competing theories can only be analogically applied to the case of technology. Clearly, a technology is not a theory, and vice-versa. However, when appropriately modified, the two concepts can be of some interest.

Running the analogy between a theory and a technology, we may say that the (pragmatic) consilience of a technology may refer to the things it enables us to do. Whereas simplicity – we argue – may refer to *the way* we are enabled to do those things. This quite broad statement may be translated into something more specific to technology: that is, simplicity can be related to the auxiliary accessories, technologies, dependencies, but also skills and knowledge one should assume to have in order to exploit all the functionalities of a given technology.

Ideally, a good technology is the one that has many functionalities and, at the same time, that does not require too many accessories, dependencies, etc.. To make a trivial example, consider two laptops. They are equipped with the same screen and keyboard size, same weight, same power, etc., but one has a battery lasting 6 hours, whereas the other runs out of battery after 3 hours. The two laptops have the same consilience, but the first is simpler, because we can go around without the need to carry the power adapter and search for an outlet to plug it in. It is more parsimonious in relation with accessories. To make another example, usability might be a sign of simplicity as well. Suppose two computer programs having the same functionalities, but one having a less user-friendly interface. In that case, the one that is easier to use is also simpler, because it does not make us think.

In the next and final section we introduce two examples in which cognitive baroqueism favors the emergence of chance-faking. The first example regards text messaging, whereas the second is about the so-called e-activism.

5 Simplicity and Consilience in Technology

Let us start with the example of text messaging. Before the advent of mobile phones appointments with friends, acquaintances, colleagues, etc., were arranged beforehand. And a delay may have caused great disappointment, since there was no way to inform about it. Mobile phones changed that. A text message allows a person to inform a friend he or she will be five minutes late. Texting may also serve to arrange a new place to meet in case of problems. Generally speaking, text messaging helps us deal with unexpected events in a unprecedented way comparing with the past. According to our definition of consilience applied to technology, mobile phone is certainly more consilient than traditional (wired) phone, as it allows us to do what we had usually done with the traditional phone, plus it allows us to stay connected with our friends or acquaintance almost wherever we are.

However, we argue that in this case consilience can be reached at the expense of simplicity. Mobile phone allows us to rely on much more flexible forms of communication with our friends and acquaintances. However, text messaging can mess things up, if misused. Indeed, when text messaging is used to inform our friend in case of an event totally independent from our will – a car crash, traffic jam, or whatever – our communication is just enhanced and enriched by a new and valuable chance, that is, flexible communication. A text may help avoid misunderstanding and possible complains due to one's delay. But, if one relies on such form of communication basically for avoiding to stick to the time (and maybe even the place) arranged for the appointment, then the chance provided by the mobile phone turns out to be counteractive, namely, a fake chance. Plus, it might create even worse misunderstanding than before. In a way one may argue that it would be much simpler to set an appointment for a time and stick to it, acting as if one did not have the chance to communicate with his or her friend.

At a more abstract level, the flexibility afforded by mobile communications takes away from the user an important communicative constraint people have usually made use of for planning and organizing their appointments. Such constraint is indeed a limitation, but it might also acquire a cognitive meaning insofar as it facilitates to keep communicating simple and therefore more reliable and just less ambiguous. Besides, the nature of such constraint – which *physically* makes mobile communication impossible – arranges one's cognitive niche in such a way that some affordances are immediately available whereas others are not. Conversely, the absence of physical constraints might literally intoxicate the cognitive niche with fake chances just like the case described above. For example, communication flexibility might afford a person to externalize her or his responsibility becoming a kind of *immoral* mediator [Magnani, 2011]. Indeed, it is an immoral mediator from a particular ethical perspective according to which one should not adopt uncooperative behaviors and thus poison relationships with friends, acquaintances, colleagues, etc..

A similar yet different example is provided by the so-called e-activism. During the last decades a number of social applications has emerged potentially promoting grass-

roots activism, political inclusion, and an alternative conception of political activism based on the massive exploitation of information technologies. The social networking website Facebook is certainly the most successful among the social softwares and platforms for connecting people and creating consensus around a common cause. There are many examples that clearly demonstrated this potential. Consider recent campaigns, such as the international mobilization to prevent Sakineh, an Iranian woman accused of adultery and murder, from being stoned to death, or the anti-FARC campaigns reported in [Neumayer and Raffl, 2008]: those campaigns are clearly devised by highly-committed and informed individuals, and then proposed unto mass public by means of mouth-to-ear dynamics (sometimes picked up and echoed by corporate news) and social networks. Astonishingly, about 100.000 people were said to confirm their participation within the Facebook group, demonstrating their solidarity with the Colombian citizens.

Sticking to the case of Facebook as representative, groups and fanpages offer the chance to get people involved in any political issue that one would like to bring up in the public sphere. Groups and fanpages allow 1) to make people's preferences visible and 2) to aggregate them so as to raise public concern about a certain issue (Sakineh, FARC, or whatever). There is not doubt that such tools are enhancing and empowering, if one looks at the chances that they potentially provide the users with. However, if one compares e-activism with traditional forms of activism, for instance, that of the 1960s, it seems that the former is characterized by superficiality and shallow connections resulting from desultory behavior [Bertolotti *et al.*, 2011]. As a matter of fact, social networks like Facebook drastically lower the bar for effective and actual political participation and involvement. However, just like in the case of text messaging, we claim that this simplification is just apparent: what actually happens is that one is just one click away from embarking on any kind of conceivable online campaign. Is that really enough for defining political participation and involvement?

If it is true that it is much easier for a person to be exposed to many more chances to get politically involved, it is also true that such easiness may just promote superficial and degraded commitment due to the number of requests one usually receives – a kind of information overloading. And so promoting actual fake chances. Analogously to the case of text messaging, here again consilience is reached at the expense of simplicity. Traditionally, activism was considered as an effect of a sincere ideological commitment, and was usually focused within a well-defined field (i.e. womens rights, civil rights, environmental issues). Conversely, the emergence of new forms of activism mediated by social networking web sites like Facebook may make paradoxically harder political activism, if one goes beyond the “I click” or/and “I like” dimension which is characteristic to social tools like Facebook. In fact, the various requests for getting involved are rather chances for a desultory form of participation characterized by a digressive and shallow, random and occasional attitude.

6 Conclusions

In this paper we have illustrated that the activity of chance-seeking might be drastically impaired by the presence of fake chances resulting from bullshitting. We have pointed out that one of the main aspects of bullshitting – considered from a cognitive perspective – is the systematic dismissal of one of the most important pillars of scientific rationality: simplicity. What is important to stress here is that the bullshitter as a chance faker fails to take into account the importance of looking at what one can assume as true and what one cannot. In the last part of the paper, we have tried to show how this attitude might affect the way people use technology. More specifically, we have illustrated how a chance – potentially benefiting for cooperation and human relations – may turn out to be a fake chance due to the adoption of a bullshitting attitude.

References

- [Abe, 2009] A. Abe. Cognitive chance discovery. In Stephanidis C., editor, *Universal Access in HCI, Part I, HCII2009 (LNCS5614)*, pages 315–323, Berlin/New York, 2009. Springer.
- [Bardone, 2011] E. Bardone. *Seeking Chances. From Biased Rationality to Distributed Cognition*. Springer, Berlin-Heidelberg, 2011.
- [Bertolotti et al., 2011] T.W. Bertolotti, E. Bardone, and L. Magnani. Perverting activism. cyberactivism and its potential failures in enhancing democratic institutions. *the International Journal of Technoethics*, 2011. forthcoming.
- [Frankfurt, 2005] H. Frankfurt. *On Bullshit*. Princeton University Press, Princeton, NJ, 2005.
- [Goffman, 1986] E. Goffman. *Frame Analysis. An Essay on the Organization of Experience*. Northeastern University Press, Boston, 1986.
- [Hample and Young, 2009] B. Hample, D. Warner and D. Young. Framing and editing interpersonal arguments. *Argumentation*, 23:21–37, 2009.
- [Hample et al., 2009] D. Hample, B. Han, and D. Payne. The aggressiveness of playful arguments. *Argumentation*, 2009. forthcoming.
- [Jackson, 2008] S. Jackson. Black box arguments. *Argumentation*, 22:437–446, 2008.
- [Magnani and Bardone, 2008] L. Magnani and E. Bardone. Sharing representations and creating chances through cognitive niche construction. The role of affordances and abduction. In S. Iwata, Y. Oshawa, S. Tsumoto, N. Zhong, Y. Shi, and L. Magnani, editors, *Communications and Discoveries from Multidisciplinary Data*, pages 3–40. Springer, Berlin, 2008.
- [Magnani, 2005] L. Magnani. Chance discovery and the disembodiment of mind. In R. Oehlmann, A. Abe, and Y. Ohsawa, editors, *Proceedings of the Workshop on Chance Discovery: from Data Interaction to Scenario Creation, International Conference on Machine Learning (ICML 2005)*, pages 53–59, 2005.
- [Magnani, 2009] L. Magnani. *Abductive Cognition. The Eco-Cognitive Dimension of Hypothetical Reasoning*. Springer, Heidelberg/Berlin, 2009.
- [Magnani, 2011] L. Magnani. *Understanding Violence. The Intertwining of Morality, Religion and Violence: A Philosophical Stance*. Springer, Berlin/Heidelberg, 2011.
- [Misak, 2008] C. Misak. Pragmatism and solidarity, bullshit, and other deformities of truth. *Midwest Studies in Philosophy*, 32:111–121, 2008.
- [Neumayer and Raffl, 2008] C. Neumayer and C. Raffl. Facebook for Global Protest: The Potential and Limits of Social Software for Grassroots Activism. In L. Stillman and G. Johanson, editors, *Proceedings of the 5th Prato Community Informatics & Development Informatics Conference*. 2008.
- [Odling-Smee et al., 2003] F. J. Odling-Smee, K. N. Laland, and M. W. Feldman. *Niche Construction. The Neglected Process in Evolution*. Princeton University Press, Princeton, NJ, 2003.
- [Ohsawa and McBurney, 2003] Y. Ohsawa and P. McBurney, editors. *Chance Discovery*. Springer, Berlin, 2003.
- [Peirce, 1931 1958] C. S. Peirce. *Collected Papers of Charles Sanders Peirce*. Harvard University Press, Cambridge, MA, 1931-1958. vols. 1-6, Hartshorne, C. and Weiss, P., eds.; vols. 7-8, Burks, A. W., ed.
- [Simon, 1993] H. A. Simon. Altruism and economics. *American Economic Review*, 83(2):157–161, 1993.
- [Thagard, 1978] P. Thagard. The best explanation: Criteria for theory choice. *The Journal of Philosophy*, 75(2):76–92, 1978.
- [Thagard, 1992] P. Thagard. *Conceptual Revolutions*. Princeton University Press, New York, 1992.
- [Thomas, 1999] H. J. Thomas. Are theories of imagery theories of imagination? An active perception approach to conscious mental content. *Cognitive Science*, 23(2):207–245, 1999.
- [Vrij, 2008] A. Vrij. *Detecting Lies and Deceit Pitfalls and Opportunities*. Wiley, New York, 2008.

Web-based Innovation Supporting System for Creative Ideas Emerging

Hao Wang, Yukio Ohsawa, Yoko Nishihara

Department of Systems Innovation, The University of Tokyo

Tokyo, Japan

wanghao@panda.sys.t.u-tokyo.ac.jp

Abstract

In this paper, we propose a novel innovation Web-based Innovation Supporting System (ISS) to improve the idea creation process and enhance the innovation capability of human. Ideas are usually stored by low frequency or zero frequency information in the dataset so that product designers are always suffering from fewer creative ideas mining. Therefore, chance discovery theory is applied in this system. The Web-based ISS is a human-centric system to enable value cognition and goes by the following process: (1) scenario graph generation, (2) customer-centric innovators market game (C²IMG) including human value cognition, value co-creation as emerging creative ideas, and creative ideas evaluation. Firstly, a mining algorithm called data crystallization is used as an extension of chance discovery for dealing with unobservable events and generating a scenario graph. Secondly, a method of C²IMG is developed to aid humans' value cognition to scenario graph and speed up creating process of innovative ideas. Finally, the analytic result of case study shows people who apply this system can create more innovative ideas than those who do not. Moreover, additional interesting patterns that indicate are discovered as well. Notably, ISS system method can be regarded as an extension of traditional data mining and machine learning for discovering more patterns with higher accuracy especially under dynamic and complex system environments.

1 Introduction

In the dynamic and competitive market, many companies have to timely launch new and creative products or services to fulfill the consumers' demands for occupying much more market share. They pay more attention on Research and Development (R&D) Department for higher technology products being created, which secure them dominate in the market. However, high technology is not necessary to create value for our society [Ueda *et al.*, 2009]. In fact, the value of products is created by dynamic interaction between consumers and product designers.

An interview and requirement meeting is a good way of data acquisition for market research and analysis, and is also an effective communication technique between product designers and customers [Carrol, 2000]. However, demand data obtained from customers is usually incomplete information. The reasons are as follows:

(1) The designers failed to understand consumers' real needs. Sometimes the consumers propose the requirements which are too ambiguous or abstract to understand for designers.

(2) The designers misunderstand consumers' real needs. The original requirements of customers are too vague.

Moreover, a creative idea as innovative knowledge for designing a new product is very important at the early stage of product development. However, ideas usually as low frequency or even zero frequency information are stored in the dataset. Thus product designers are always suffering from mining creative ideas for new product and service development by traditional data mining and statistics methods.

In the last few years, chance discovery has been widely applied in various research areas, especially in business area. Chance discovery is a human-computer interaction process to detect rare but important chances for decision making. A chance in chance discovery means to understand an unnoticed event or situation which can be uncertain but significant for a decision by human beings [Ohsawa and McBurney, 2003]. KeyGraph is a visualization tool with the KeyGraph algorithm which can generate scenario graph for aiding human's value cognition in the process of chance discovery [Ohsawa *et al.*, 1998]. In the case of marketing, users of KeyGraph sold out new products, and made real business profits [Ohsawa and Fukuda, 2002]. However, the researchers gradually recognized a new problem in chance discovery where KeyGraph failed to visualize important events which are neither visible nor observed as data [Maeno and Ohsawa, 2007]. Therefore, Ohsawa [Ohsawa, 2005] proposed a breaking-through method named data crystallization where dummy nodes corresponding to the unobservable events are inserted into the basket data, and then are visualized by KeyGraph. Horie, Maeno and Ohsawa [Horie *et al.*, 2007] applied the method of data crystallization for designing new products in a real company, the results of which illustrated its effect to industrial decision making. Wang, Ohsawa and Nishihara [Wang *et al.*, 2011] proposed a system method of

innovative knowledge elicitation based on chance discovery for new product design.

Even so, there are fewer approaches proposed under a systematic or united framework for the purpose of innovative product design from knowledge elicitation to evaluate knowledge representation. How to develop a systematic method, which can capture important but low or even zero frequency data, and combine visual technologies with the methods of enhancing cognition capability of human, to support human in creating innovative ideas as future chances, has become an important research topic.

In this paper, a framework of Web-based ISS is introduced in section 2. Section 3 mainly presents the scenario graph generation process and related algorithm by chance discovery with data crystallization. For better understanding scenario graph to aid humans' innovation, customer-centric innovators market game (C²IMG) is developed in section 4. Experiments with C²IMG are presented in section 5. The discussion and conclusion is in the final section.

2 Web-based Innovation Supporting System

A Web-based ISS is proposed and the interface of it is shown in Figure 1, which includes three main functional modules: (1) Scenario Generation and Innovation Module, (2) Requirement Releasing Module, and (3) Communication Module. These three modules will be introduced in C²IMG section.

In this system, a scenario graph for human's value cognition is generated by an algorithm of data crystallization based on chance discovery theory. The scenario graph shows the topological relations between nodes. Based on the scenario graph as a game board, a method of Customer-centric Innovators Market Game (C²IMG) is developed to aid humans' value cognition for scenario graph and speed up emerging process of creative ideas.

Web-based ISS supplies a dynamic interaction platform between consumers and product designers for creative ideas emerging in the circumstance of C²IMG. In the following section, it will mainly introduce scenario generation method and how C²IMG improve humans' value cognition ability to understand the scenario graph and promotes humans' innovation.

3 Scenario graph generated by chance discovery with data crystallization

Chance discovery is a human-computer interaction process. For assisting the process of chance discovery, Keygraph as a tool of creating scenario graph is developed for human's cognition for chances.

3.1 KeyGraph: a visualization tool for scenario graph generation

The visualization graph is shown in Figure 1. Suppose that data D is stored in a document. The data format is as below:

D = Item1, Item2, Item3.
Item1, Item2, Item3, Item4.

Item3, Item4, Item5, Item6.

The main steps of Keygraph algorithm are outlined as follows.

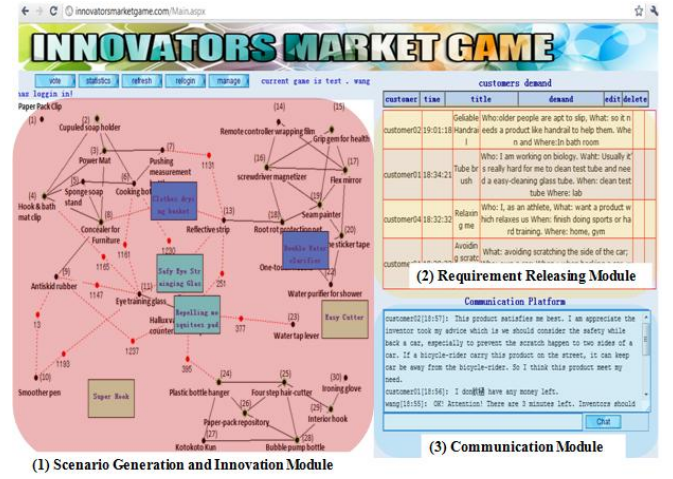


Figure 1: The interface of Web-based ISS

Step 1: Obtain the clusters composed of co-occurrence high frequency items.

- Extracting high-frequency items. Items in the data are sorted by frequency of occurrence. These high-frequency items are represented as black nodes in the graph G.
- Extracting high-frequency item-pairs. High-frequency item-pairs means item-pairs occurring in the same line of data have high co-occurrence frequency. Item-pairs are also sorted by frequency of occurrence. High-frequency item-pairs are linked with black solid line in the graph G, thereby forming clusters. A measure of co-occurrence of item-pairs is defined as the Jaccard co-efficient

$$J(I_i, I_j) = \frac{Freq(I_i \cap I_j)}{Freq(I_i \cup I_j)} \quad (1)$$

where, I_i and I_j are the high-frequency items. $Freq(I_i \cap I_j)$ is the probability of item I_i and I_j co-occurring in the same line. $Freq(I_i \cup I_j)$ is the probability of either item I_i or I_j occurring in the line.

Step 2: Obtain chance candidates. The items strongly co-occurring with clusters are obtained as chance candidates.

- Extracting high-key items. The measure of the tightness between an item I and a cluster C is defined:

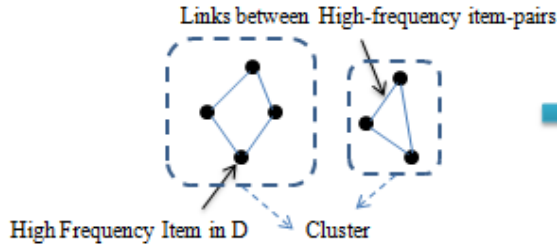
$$Key(I) = 1 - \prod_{C \in G} [1 - J(I, C)] \quad (2)$$

The key value of each item in the data is calculated by Eq. (2). The high-key items are added as nodes with red color if they are not already in the graph.

- Extracting chance links. According to the Eq. (1), each high-frequency item and each high-key item is calculated. Links connecting to high-frequency are

sorted by their J values for each high-frequency

Step 1: Obtain the clusters composed of co-occurrence high frequency items.



item, and are added in the graph with red dot lines.

Step 2: Obtain chance candidates.

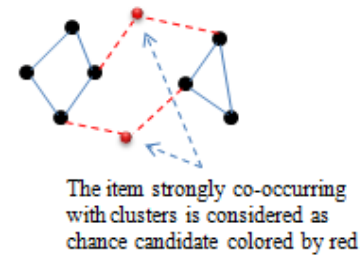


Figure 2: The Generating Procedure of Scenario Graph

In chance discovery, human make consciousness of the visual graph made by Keygraph. The clusters obtained imply a basic context. The red nodes linking different clusters are considered as candidates of chance that can be important for decision making in the future.

3.2 Data crystallization: chance discovery extended for dealing with unobservable events

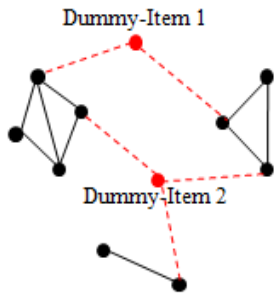


Figure 3: A graph visualized by Keygraph based on the data D' including the dummy-items

For incomplete or ill-structured data, the methods introduced above in chance discovery cannot deal with this kind of problems well. Some important events probably don't exist in the data. Therefore, how to detect unobservable significant events is desired to solve. Data crystallization is proposed to present the hidden structure among events including unobservable events.

The process of data crystallization is as follows:

Step 1: Inserting the dummy-item. Given the data D , a dummy-item is inserted into each line of the data D , obtaining a new data D' . If there are the same dataset in two or more line, these lines should be inserted the same dummy-item.

```
D' = Item1, Item2, Item3, Clone-Item1
      Item1, Item2, Item3, Item4, Clone-Item2
      Item3, Item4, Item5, Item6, Clone-Item3
      .....
      .....
```

Step 2: Visualizing the data D' by Keygraph, see Figure 2. Dummy-items appear between the clusters.

Step 3: Eliminating meaningless dummy-items. The dummy-items which don't appear in the graph and connect different clusters as a bridge are deleted from D' . For example, dummy-item3 does not show in Figure 2, thus the dummy-item3 should be deleted in the data D' .

Finally, the following will be obtained:

(1) A new data set with dummy-item, corresponding to hidden events that connect substructures in each level.

(2) $\text{KeyGraph}(D, M_1, M_2, M_3)$ for the given data D and arbitrarily determined values of M_1, M_2 and M_3 . By increasing M_2 , the output to the higher level of the hidden structure can be obtained. By decreasing M_2 , the granularity of the visualized structure is increased.

A tool called Data Crystallizer which is developed for inserting dummy items to the given data runs the Algorithm 1 below:

Algorithm 1 The procedure of data crystallization

Input: $k:=1, M_1:=$ a value provide by user
Output: A new data set D' .

```
1: for  $M_2 = 1$  to  $M_1(M_1+1)/2$  do
2:   for all  $i, j \in 0, 1, \dots, N$  such that  $j \geq i$  do
3:     if  $\text{line}_i$  and  $\text{line}_j$  are equal then
4:       insert  $(D, k, i, j)$ ;
5:     end if
6:   end for
7:    $H := \text{keygraph}(D, M_1, M_2, M_3 := M_1/2)$ ;
8:   for  $j=1$  to  $N$  do
9:     if  $j \notin H$  then
10:      delete  $(D, k, j)$ ;
11:     end if
12:   end for
13:   if  $H \neq \text{Hidden}_k$  then
14:      $k := k+1$ ;
15:      $\text{Hidden}_k := H$ ;
16:   for  $m = 0$  to  $(k-1)$  do
```

```

17:      delete (D, m, Hidden_m ∩ H);
18:      Hidden_m := Hidden_m \ H;
19:    end for
20:  end if
21: end for

```

where D is the basket data to be analyzed with Keygraph in the function of KeyGraph (D, M_1, M_2, M_3). M_1, M_2 and M_3 separately indicates the numbers of black nodes, black lines and hubs. N is the number of lines (co-occurrence units) in the data, and $line_j$ represents the set of items in the j -th line. H represents the set of line-numbers where dummy items appearing on the bridges of the current KeygGraph are positioned in the data. $Hidden_i$ means the set of line-numbers with a dummy item in the i -th level. The function insert (D, k, i, j) means to insert k_j , the dummy node for the j -th line in the k -th level of crystallization, to i -th line of data D and from data D . delete (D, k, j) means to delete k_j , the dummy item for the j -th line on the k -th level, for all its appearance in data D .

4 Customer-centric innovators market game

In last section, although a scenario graph is obtained, it's a little hard for humans to understand the scenario only depending on individual knowledge. Therefore, we developed a new method to aid humans' value cognition for the scenario.

In customer-centric innovators market game (C²IMG), there are also three same roles: dealer, customer and innovator, in which the core role is customer. There are also three kinds of innovators representing three different competing companies. As shown in Figure 1, Module (2) is specified for the customer to propose new demands; Module (1) is used to generate creative ideas by innovators; Module (3) as a communication platform enables all the players to share, merge and evaluate their knowledge.

Each player needs to be clear that

- **idea cards** corresponding to each black nodes as basic idea source for all the players. Each idea card records the introduction and function of each current product in the market;
- **game board** which is a scenario graph showing idea cards' relationship by KeyGraph with data crystallization;
- **virtual currency** for customers to buy new or creative products from innovators;
- **Voting** to those customers whose demands are valuable for creating new idea by innovators.

When the game starts, each customer has fixed virtual currency for buying new products created by innovators, and each innovator can vote to customers for supplying valuable requirements and opinions.

The game is played by the following process:

Step 1: Customers are required to propose new requirements according to these basic cards and scenario graph, and meanwhile write them down in Module (2). Each record should be written with one need. After finishing writing,

customers are also required to announce their requirements to innovators in Module (3). The requirement record should be written down by customers in accordance with the following format, consisting of What, Who, When and Where, shown in Figure 4. We call it 4W requirement. "What" needs customers to propose real requirement or complaint. "Who" refers to the requirement proposed by which kind of customer group. "When and Where" refers to the time and circumstance of the requirement.

Step 2: Innovators try to create new products by combining existing products (basic cards) and customers' requirements, and record new ideas (products) in Module (1) and are encouraged to create new products which correspond to the red nodes in the game board. As shown in Figure 5, innovators should write creative idea down by How and Why. When an innovator adopts requirement of a customer and creates new product, the innovator should award a product certificate to the customer for his or her valuable requirement.

Step 3: Repeat Step 1 and 2, consumers and innovators communicate with each other in Module (3) for co-creating new products. The game will last 1-2 hours as required.

When the game is over, the innovator in three who has the biggest maximum score by Eq. (3) will be the winner in all of innovators and be honored with "Best Innovators". Similarly, the customer who achieved the most votes from innovators will win the honor of "Best Customer".

$$Score(i) = T_i / N_i \quad (3)$$

where T_i indicates total profit of the i th innovator; N_i is the number of ideas created by the i th innovator.

Need Number & Name: (Customer)

What:

Who:

When and where:

Figure 4: A sample of 4W requirement from customer

Idea Number & Name: (Innovators)

How:

Why:

Figure 5: A sample of H-W creative idea proposed by innovator

5 Experiment using C²IMG

The C²IMG of 7 players was hosted in the Web-based ISS. This game required an hour to complete.

The final result of C²IMG is shown in Table 1 and Table 2. Innovator3 had the highest score and won the honor of “Best Innovator”, while “Best Customer” was awarded to Customer2.

Innovator No.	Total Profit	The number of ideas	Score	Rank
1	6	6	1	2
2	5	5	1	2
3	9	6	1.5	1

Table 1 The final result of C²IMG for innovators

Customer No.	The number of product certificates	Rank
1	3	2
2	4	1
3	2	3

Table 2 The final result of C²IMG for customers

5 Discussion and Conclusion

Through experiment above, the following important and useful patterns are obtained.

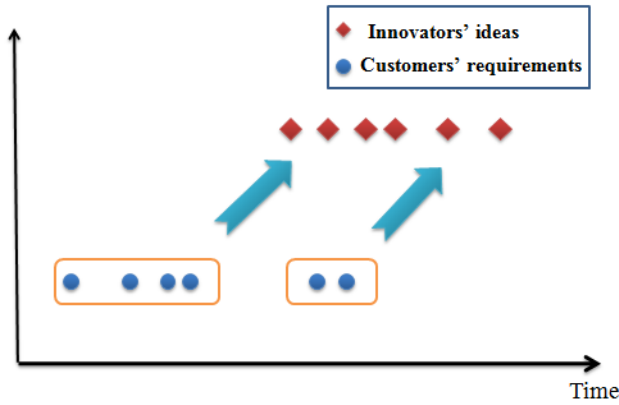


Figure 6: The relationship of ideas from innovators and requirements from customers

1. Customers' needs drive innovators and market innovation. Figure 6 shows it always has the pattern that customers firstly propose new requirements after the game starts, and then innovators create new ideas by combination according to these requirements.

2. It indicates that the number of proposed requirements is almost equal to the number of created ideas. Therefore, based on scenario graph generated by Keygraph, how to capture more and more real voice of customers effectively is still our future work, if we want to gain a growing number of creative ideas in the finite time.

3. Humans who apply this system method can create more creative ideas than those who do not.

We had successfully organized a table Innovators Market Game (IMG) in a special session of the 11th International Symposium on Knowledge and Systems Sciences held in Xi'an, China [Wang *et al.*, 2011]. A comparison is made between Table IMG and Web ISS.

As shown in Figure 7, Web ISS gained requirements and creative ideas as more than two times many as Table IMG.

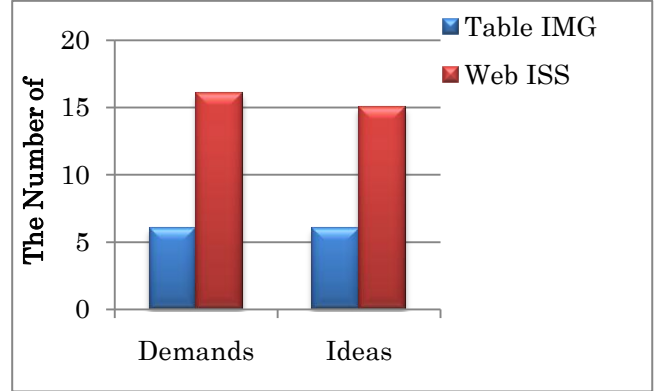


Figure 7: The result comparison when using Web-based ISS and not

4. ISS system method can be regarded as extension of traditional data mining and machine learning. The principle work of Web-based ISS is shown in Figure 8. New value as the output of the system is added into the original data file for further mining, which visualizes meaningful patterns happen in future. However, traditional methods of data mining and machine learning normally are based on historical data selected to discover interesting patterns. Thus it cannot accurately predict what will happen, especially in dynamic and complex environments. Furthermore, creative ideas as latent information cannot be easily captured by most of data mining algorithms because of their low frequency or even zero frequency character in the database.

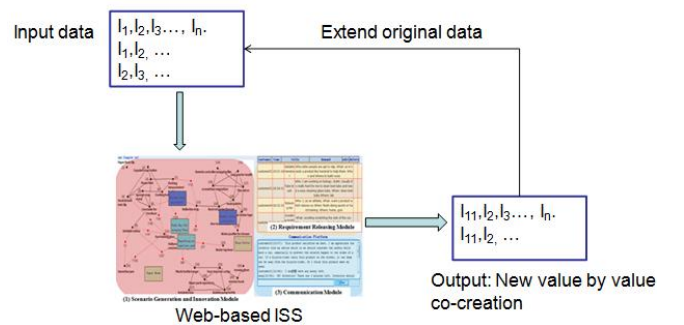


Figure 8: The principle work of Web-based ISS

Therefore, this paper puts forward a novel Web-based Innovation Supporting System (ISS) for human's innovation, and also presents an effective solution named C²IMG to

discover ideas based on the generated scenario graph, especially creative ideas for new product or service development. In C²IMG, on the principle of customer-centered, various requirements from the consumers are selected. According to the consumer requirements, the product designers based on their working experience create new ideas for breakthrough products. The experiment results showed some interested patterns, of which humans who applied this system method can make more creative ideas emerging than those who did not. This system can also be applied as Intelligent Tutoring Systems (ITS) to meet its new challenges and directions (Conati, 2009).

Acknowledgements

The author was supported through the Global COE Program “Global Center of Excellence for Mechanical Systems Innovation,” by the Ministry of Education, Culture, Sport, Science and Technology.

References

- [Carrol, 2000] J. M. Carrol. *Making Use Scenario-based design of Human-computer interactions*. The MIT press, Cambridge, Massachusetts, 2000.
- [Conati, 2009] C. Conati. Intelligent Tutoring Systems: New Challenges and Directions. In *Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI'09)*, 2-7.
- [Horie *et al.*, 2007] K. Horie, Y. Maeno and Y. Ohsawa. Data crystallization applied for designing new products. *Journal of Systems Science and Systems Engineering*, 16(1):34-49, 2007.
- [Maeno and Ohsawa, 2007] Y. Maeno and Y. Ohsawa. Human – Computer Interactive Annealing for Discovering Invisible Dark Events, *IEEE Transactions on Industrial Electronics*, 54(2):1184 -1192, 2007.
- [Nakamura and Ohsawa, 2009] J. Nakamura and Y. Ohsawa. A shift of mind - Introducing a concept creation model. *Information Sciences*, 179 (11):1639-1646, 2009.
- [Ohsawa, 2005] Y. Ohsawa. Data crystallization: chance discovery extended for dealing with unobservable events, *New mathematics and natural science*, 1(3):373-392, 2005.
- [Ohsawa and Fukuda, 2002] Y. Ohsawa and H. Fukuda. Chance Discovery by Stimulated Group of People, An Application to Understanding Rare Consumption of Food. *Journal of Contingencies and Crisis Management*, 10(3): 129–138, 2002.
- [Ohsawa and McBurney, 2003] Y. Ohsawa and P. McBurney (eds). *Chance Discovery*, Springer, Heidelberg, 2003.
- [Ohsawa *et al.*, 1998] Y. Ohsawa, N. E. Benson, and M. Yachida. KeyGraph: Automatic Indexing by Cooccurrence Graph Based on Building Construction Metaphor. In *Proceedings of the Advanced Digital Library Conference (ADL'98)*, pages 12–18, 1998.
- [Ueda *et al.*, 2009] K. Ueda, N. Nishino, and T. Takenaka. Producer decision-making in market with network externalities. *CIRP Annals - Manufacturing Technology*, 58: 413–416, 2009.
- [Wang *et al.*, 2010] H. Wang and Y. Ohsawa, Y. Nishihara and L. Guo. An effective method to discover emerging chance in chance discovery. In *Proceedings of 9th IEEE international conference on cognitive informatics (ICCI 2010)*, pages 639 – 644, 2010.
- [Wang *et al.*, 2011] H. Wang and Y. Ohsawa, Y. Nishihara. A System Method to Elicit Innovative Knowledge Based on Chance Discovery for Innovative Product Design. In *International Journal of Knowledge and Systems Science*. (In press)

Discovering Probabilistic Models from ADS-B Data Sets using Digital Pheromones

Kirk Ogaard

Computer Science Department
University of North Dakota
Grand Forks, ND, USA
ogaard@cs.und.edu

Ronald Marsh, Ph.D.

Computer Science Department
University of North Dakota
Grand Forks, ND, USA
rmarsh@cs.und.edu

Abstract

The University of North Dakota is developing airspace within the state where Unmanned Aircraft Systems (UASs) can be flown without an onboard sense and avoid system or Temporary Flight Restrictions (TFRs). With funding from the U.S. Air Force, a mobile ground-based system capable of detecting cooperative aircraft flying in the surrounding airspace, and the software to display information about those aircraft to UAS operators, was developed. An ongoing area of research for this system is its risk mitigation component. The risk mitigation component will be responsible for calculating the total collision risk for an Unmanned Aircraft (UA) flying in a specific Class E airspace configuration. If accurate probabilistic models for pilot behavior are determined, these models could be used in the implementation of the risk mitigation component. In this paper the authors present the results of subpath discovery and data mining of an Automatic Dependent Surveillance – Broadcast (ADS-B) data set from a 3 month period in 2010. Arbitrarily complex subpaths were automatically extracted from the ADS-B data set using a digital pheromone algorithm. Then, probabilistic models of pilot behavior were data mined from this set of discovered subpaths using a variant of the Genetic K-Means Algorithm (GKA).

1 Introduction

Unmanned Aircraft (UAs) have many unique advantages over manned aircraft. With no pilot on board, UAs can be used in dangerous situations or for routine and mundane operations. With no pilot to account for in their design, UAs can be designed to carry more fuel and payload, thus expanding their operational capabilities. With its smaller size and weight, a UA also benefits from lower manufacturing and operating costs.

However, flying UAs in the U.S. National Airspace (NAS) can be problematic as it is unclear whether or not Federal Aviation Regulations even apply to unmanned aircraft [Weibel and Hansman, 2005]. There are also safety guidelines that Unmanned Aircraft System (UAS) operators must contend with. A previous analysis of ground and midair

collision risks found that most UAS operations would not meet the target levels of safety established by the Federal Aviation Administration (FAA) without the incorporation of a risk mitigation strategy.

Given the regulatory difficulties of flying UAs in the NAS, the University of North Dakota (UND) is identifying airspace within the state where organizations interested in developing UASs can test and operate their systems without needing onboard sense and avoid capabilities. UND hopes to provide an interim Gated Phased Array Radar – Risk Mitigation System (GPAR-RMS) strategy to allow UAS research and development outside restricted airspace. This will facilitate the development of appropriate regulations for UAS operations.

GPAR-RMS, which is currently under development, is intended to be an extension of the ground-based observer. GPAR-RMS integrates aircraft telemetry data from sensor systems, including an Automatic Dependent Surveillance – Broadcast (ADS-B) receiver and ground stations receiving telemetry data from Global Positioning System (GPS) equipped UAs, and displays those data on two types of Information Display Systems (IDSs)—the Range Control Center (RCC) IDS and the Ground Observer (GO) IDS [Marsh *et al.*, 2011]. An ongoing task is the development of the risk mitigation component for GPAR-RMS. The current risk mitigation component is only a simple prototype created to facilitate the testing of other components of GPAR-RMS. We are currently developing a more advanced risk mitigation component to replace this prototype.

The problem of calculating the total collision risk for a UA flying in a specific airspace configuration is algorithmically complex. Although the set of basic maneuvers that a pilot can perform with a manned aircraft is known, the flight path a pilot chooses for a manned aircraft can be composed of any combination of these basic maneuvers. Furthermore, according to FAA regulations [Federal Aviation Administration, 2010], pilots flying manned aircraft under Visual Flight Rules (VFRs) in Class E airspace (i.e. at altitudes below 18,000 feet MSL) are not required to be under positive control by Air Traffic Control (ATC). Hence, the probable flight paths taken by manned aircraft flying under VFRs in Class E airspace remain unknown.

Our previous approach to this problem [Marsh and Ogaard, 2010] was to classify pilot maneuvers extracted from ADS-B data sets according to a static classification

scheme of basic maneuvers typically performed by pilots. Once classified, these subpaths were then data mined using a variant of the Genetic K-Means Algorithm (GKA) [Krishna and Murty, 1999; Lu, *et al.*, 2004].

Data mining ADS-B data sets (using algorithms such as GKA) can aid in discovering probabilistic models of pilot behavior in manned aircraft. An ADS-B data set contains very accurate data concerning the flight paths of ADS-B-equipped aircraft over a specific time period. Accurate estimations of the probabilities of frequently occurring maneuvers could be mined from a massive ADS-B data set. These estimated probabilities could then be used to fully implement the risk mitigation component of GPAR-RMS.

Although research has been done on algorithms for mining movement patterns from 2D GPS-based data sets [McCall, *et al.*, 2007] and predicting movement patterns by correlating GPS data to other variables [Taniar and Goh, 2007], these algorithms are only capable of discovering basic movement patterns in 2D GPS-based data sets. The more advanced approach described in this paper efficiently and automatically discovers arbitrarily complex subpaths in ADS-B data sets using digital pheromones. The subpaths thus discovered can then be data mined to find correlations with other variables, such as an aircraft's altitude and its proximity to the nearest private airport.

2 Subpath Discovery

The probability of a collision is based on the proximity of the aircraft to each other, the performance characteristics of the aircraft (which are determined by the type of aircraft), and the probabilities of the pilots of those aircraft performing various types of maneuvers. Since pilots flying aircraft near commercial airports are under positive control by ATC, their flight paths are deterministic. Pilots flying aircraft that are not under positive control by ATC (e.g. pilots flying under VFRs in Class E airspace) have nondeterministic flight paths that are much more difficult to predict. Thus, for aircraft flying in Class E airspace, the aircraft's proximity to private airports, rather than commercial airports, was considered more likely to influence the types of maneuvers the pilot performed. The data mining software we developed, the ADS-B Data Miner (ADM), was used to mine ADS-B data sets in order to extract probabilistic models of pilot behavior as a function of the aircraft's type (e.g. an Airbus 320), altitude, and proximity to the nearest private airport. The largest ADS-B data set we mined with ADM consists of ADS-B data received by base stations around the world during a 3 month period in 2010.

Although the ADS-B protocol is the only data format used by ADS-B-equipped aircraft to transmit their real-time telemetry data to other ADS-B transceivers (onboard ADS-B-equipped aircraft or located at base stations), several different data formats are used to store those data for later analysis. The three data formats we have encountered are the Comma-Separated Values (CSV) format, the Extensible Markup Language (XML) format, and the raw binary format used by the Garmin GDL 90 ADS-B transceiver to send data to display devices. The data sets we mined with ADM consist of a large number of text files (in CSV or XML format) containing the raw data from ADS-B transmissions from many different aircraft over an extended period of time.

The data contained in those ADS-B transmissions were stored in the text files in the same chronological order that the base station received them. Thus, if an aircraft flew for many hours and/or conducted many flights, the ADS-B data for the aircraft are distributed amongst many different text files.

Most of the data contained in ADS-B transmissions are not relevant to the analysis of the types of maneuvers performed by pilots during their flights. Also, the data streams broadcast from ADS-B transceivers are time-ordered sequences of discrete 3D GPS positions held by ADS-B-equipped aircraft, whereas the normalized flight paths of the aircraft (rather than the discrete positions along those flight paths) are more important to the analysis of pilot maneuvers. Thus, it is necessary to extensively preprocess the raw ADS-B data, which are stored in heterogeneous data formats, into a more useful, homogeneous format prior to any data mining.

ADM performs all of its data preprocessing via SQL commands that operate on tables in a relational database (see figure 1). ADM performs five phases of data preprocessing on the ADS-B data for each aircraft type in the data set: 1) extracting the ADS-B data from the text files and inserting the relevant data into a relational database, 2) normalizing the data in the database, 3) constructing normalized vectors from the discrete 3D GPS positions in the database, 4) discovering digital pheromone trails from subpaths that are common to multiple normalized flight paths, and 5) dynamically segmenting the normalized flight paths into subpaths using those digital pheromone trails. Completing these five phases of data preprocessing is necessary before any data mining for probabilistic models of pilot behavior can occur.

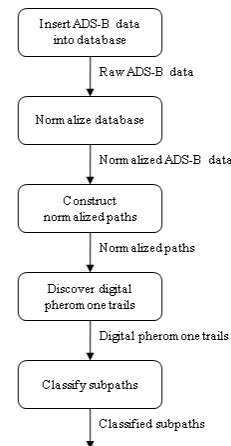


Figure 1. The flow chart for the five phases of data preprocessing.

2.1 Constructing Normalized Paths

The first phase of data preprocessing involves reading the raw ADS-B data obtained from the base stations, discarding irrelevant data, insuring each aircraft has a unique numeric identifier so it is easily tracked, and inserting the relevant data from each aircraft into the relational database. During this phase, ADM retains the following fields from the ADS-B data: a) the International Civil Aviation Organization (ICAO) identifier (if available), b) the type of aircraft, c) the tail number, d) the latitude, longitude, and altitude, e) the horizontal and vertical velocities, and f) the time of reception. The ICAO identifier is useful for data mining

because it is a number that uniquely identifies an ADS-B-equipped aircraft. However, ICAO identifiers were missing from the largest ADS-B data set we mined. Thus, ADM generated a unique number for each aircraft in this ADS-B data set based on its tail number.

The second phase involves normalizing the data in the database. Redundant data, e.g. an aircraft maintaining the same GPS position for several seconds, are discarded. Data for all flights outside of NAS and any flight segments outside of Class E airspace (i.e. above 18,000 feet MSL) are also discarded. Since an aircraft's heading is represented as a compass direction by the ADS-B protocol, it is necessary during preprocessing to normalize the aircraft headings obtained from the ADS-B data. The headings are normalized with respect to the heading currently considered the straight-flying direction for the aircraft. If the aircraft flew with the same heading h for two or more consecutive data points, then h would be considered its current straight-flying direction. Thus, to normalize the straight-flying heading h , it is rotated by $90-h$ degrees so that it is normalized to an angle of 90° . Likewise, all other aircraft headings are rotated by the same amount until the next change to the straight-flying direction. Since ADS-B data do not explicitly specify the angles of ascent for aircraft, these fields are calculated from the aircraft's horizontal and vertical velocities.

The third phase involves constructing vectors from the discrete 3D GPS positions for aircraft in the database. Consecutive ADS-B data points with the same heading and ascent angles are merged to form the longest possible vectors. The magnitude of these vectors is measured in time, not distance, because while aircraft fly at different speeds (and thus cover different distances in the same amount of time), it is reasonable to assume that the pilots of those aircraft require about the same amount of time to perform the same types of maneuvers.

2.2 Discovering Digital Pheromone Trails

The fourth phase involves the discovery of digital pheromone trails in the normalized flight paths for the aircraft. Digital pheromone algorithms have been successfully applied to many optimization problems [Parunak, *et al.*, 2002; Ando, *et al.*, 2006; Chu, *et al.*, 2007; Narzt, *et al.*, 2007; Kalivarapu, *et al.*, 2008; Kannampallil and Fu, 2009]. These algorithms are based on how swarming insects interact with their environment.

Swarming insects use pheromones to mark important parts of their environment. Ants are one example of such insects. Ants deposit pheromone trails while they randomly wander around in search of food. Ants use two types of pheromones for this behavior—a food pheromone and a nest pheromone. When an ant finds food it carries the food back to its ant hill. The ant deposits a trail of food pheromones while following a trail of nest pheromones previously deposited by other ants. When an ant departs its ant hill in search of food it deposits a trail of nest pheromones while following a trail of food pheromones. Pheromones from overlapping trails will reinforce each other. Thus, if an ant follows the closest trail with the strongest food pheromones, it will have the best chance of finding food. Conversely, if ants do not continuously reinforce pheromone trails by following them, the trails will slowly evaporate. Finally, pheromones from nearby trails will diffuse to surrounding

areas. Thus, ants that are near pheromone trails can still detect them.

In order to dynamically extract classification schemes for the types of maneuvers frequently performed by pilots, we represent each normalized flight path as a separate digital pheromone trail deposited by the aircraft. If every normalized flight path is compared with every other normalized flight path, then subpaths will be discovered that are common to several paths. These common subpaths are the areas where digital pheromones from normalized flight paths are reinforcing each other (see figure 2). For example, if a subpath is common to two normalized flight paths, its corresponding digital pheromone trail will have a strength of 2. Digital pheromone trails with greater strengths are thus more likely to represent actual maneuvers performed by pilots of manned aircraft.

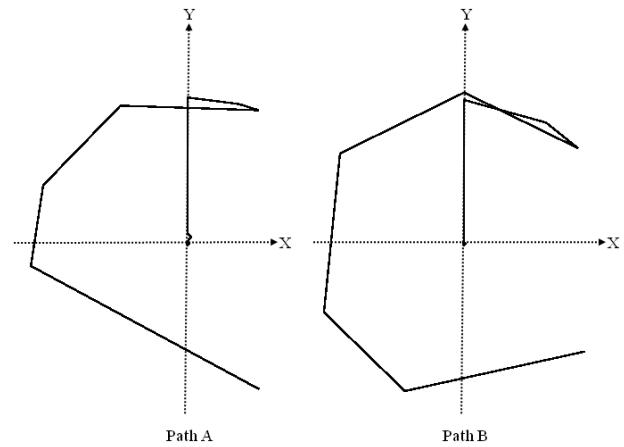


Figure 2. Normalized paths (projected in 2D) with one common subpath. The common subpath has a strength of 2.

During digital pheromone trail discovery, the shapes of the digital pheromone trails are important, not the GPS positions of their endpoints. Thus, each digital pheromone trail starts at the same point in the coordinate system. Also, normalized flight paths are represented internally as ordered sequences of unit vectors to facilitate comparisons. Every normalized flight path for a given aircraft type is compared with every other normalized flight path for that aircraft type on a vector by vector basis. If two vectors are within some constant diffusion distance d of each other, then they are considered a match. Since only the distances between unit vectors will be calculated, a value must be used for d that is less than 2. Any value for d that is greater than or equal to 2 will result in a case where every vector is within diffusion distance of every other vector. Also, any digital pheromone trails with lengths of less than 5 are discarded, because it is unlikely that the pilot of a manned aircraft could complete a standard maneuver in less than 5 seconds.

The algorithm we designed for digital pheromone trail discovery performs many distance calculations. To improve its run-time efficiency, a vector proximity map is constructed. A vector proximity map is a 2D Boolean array. It can determine if any two unit vectors are within some diffusion distance d of each other in constant run-time. The angles (i.e. heading and ascent angles) for each of the vectors are encoded as integers. These encoded integers are then used as indexes into the vector proximity map. The

element in the vector proximity map corresponding to those two vectors is 1 if the vectors are within diffusion distance of each other (i.e. the vectors match), or 0 otherwise.

The strength of a digital pheromone trail is a potential indicator of how frequently pilots performed this type of maneuver. In digital pheromone algorithms, the strengths of digital pheromone trails decrease over time through evaporation unless they are continually reinforced by new digital pheromones. However, the time elapsed according to the system clock is not a useful control variable for our problem, since ADS-B data do not arrive at the base station at a constant rate. The rate at which the ADS-B data arrive at the base station is directly dependent on the amount of air traffic currently in the vicinity of the base station. Thus, a better control variable for our problem is the amount of new ADS-B data that have arrived since the last update to the database.

Future work will include the development of an incremental algorithm for digital pheromone trail discovery that evaporates existing digital pheromone trails based on the amount of new ADS-B data that arrive at the base station. Thus, unless new ADS-B data arriving at the base station continually reinforce the existing digital pheromone trails stored in the database, they will eventually evaporate, replaced by new digital pheromone trails. The discovery of digital pheromone trails thus results in a set of frequently occurring subpaths in the ADS-B data for a specific aircraft type (see figure 3).

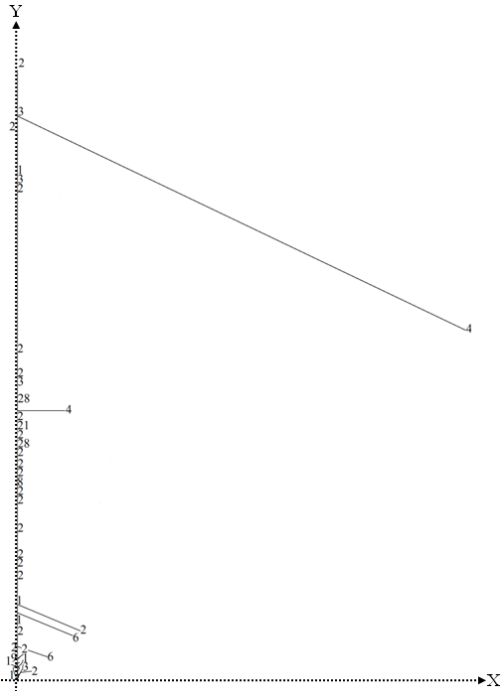


Figure 3. The digital pheromone trails (projected in 2D) discovered for the Gulfstream IV and their corresponding strengths.

2.3 Classifying Subpaths

The fifth and final phase of data preprocessing segments the normalized flight paths into subpaths using the digital pheromone trails discovered by the previous phase. The subpaths identified during this phase are stored in the

database for later retrieval during data mining. Each normalized flight path P is segmented into subpaths in an iterative manner, starting with its first vector, using the following algorithm. The current vector v_p from P is compared with the first vector v_t from every digital pheromone trail in the database whose length is less than or equal to the length of P . If the vectors v_p and v_t are within some constant diffusion distance d of each other, then the digital pheromone trail is a potential match. If the vectors v_p and v_t are not within diffusion distance d of each other, then the digital pheromone trail is not a potential match.

Next, the algorithm determines if any of the potentially matching digital pheromone trails are complete matches to any subpaths in P that start with the current vector v_p . This is determined by comparing each potential match with a subpath of the same length in P that starts with the current vector. From the set of complete matches to subpaths in P , the digital pheromone trail with the greatest strength is selected. If there are multiple such digital pheromone trails, the longest digital pheromone trail from the set of strongest matches is selected. The matching digital pheromone trail, having some length l_p , will thus match some subpath from P of length l_t . The next subpath segmented from P will start immediately after the end of the previous subpath. If there are only a few normalized flight paths for a particular aircraft type, it is possible for a normalized flight path to have subpaths that are not common to any other normalized flight paths. Such subpaths are defined by exclusion and cannot be classified.

The set of digital pheromone trails discovered for a particular aircraft type thus form a dynamic classification scheme for the types of maneuvers a pilot is likely to perform when flying that type of aircraft. Each digital pheromone trail is potentially a unique type of maneuver intentionally performed by pilots. In some cases, there are may be two or more digital pheromone trails with different strengths that match the same subpath in a normalized flight path. In these cases the digital pheromone trail with the greatest strength (i.e. the one that occurs most frequently in the data) is always used. However, since we do not know the actual intentions of the pilot, it is possible (though unlikely) that the pilot actually performed a maneuver corresponding to a digital pheromone trail of lesser strength.

There may have been other factors affecting the aircraft's heading, ascent angle, and altitude than just the pilot's control inputs. Also, the uncertainties in the aircraft's horizontal and vertical position (inherent to the ADS-B protocol) introduce a measurable amount of error. To correct for these anomalies, ADM rounds up the aircraft's heading and ascent angles to the nearest multiple of 2 degrees, and the aircraft's altitude to the nearest multiple of 2 meters. Thus, we consider the values for the heading, ascent angle, and altitude of the aircraft to be accurate indicators of the pilot's intent to within ± 1 degree for headings and ascent angles, and to within ± 1 meter for altitudes. This is a valid assumption in most cases, since, for example, the standard rate of turn for a pilot is 2 degrees/second, which means pilots won't normally turn their aircraft at a rate of less than 2 degrees/second.

The aircraft's altitude, proximity to the nearest commercial airport, and proximity to the nearest private airport at the start of each subpath are calculated and/or

stored in the database for later retrieval. The proximity to the nearest commercial or private airport is the geodesic distance from the aircraft's latitude/longitude position to the latitude/longitude position of the nearest commercial or private airport, respectively. We calculated proximity to airports by searching an airport database for commercial or private airports, respectively, which were near the aircraft (using the search algorithm described in [Marsh and Ogaard, 2010]), finding the distance from the aircraft to each of the nearby commercial or private airports, respectively, and then selecting the minimum of those distances.

Although the volume of controlled airspace around commercial airports varies from airport to airport, the maximum volume of controlled airspace around any airport in NAS is 27,780 meters horizontally and 3,000 meters vertically. Thus, if an aircraft was within this maximum volume of controlled airspace around some airport at the start of a subpath, the subpath is considered to be within controlled airspace. It is therefore excluded from further consideration. However, if the aircraft's subpath started in uncontrolled airspace, and then later veered into controlled airspace, it is not excluded.

3 Data Mining Subpaths

ADM performs all of its data mining via SQL commands that operate on tables in a relational database. ADM performs its data mining in two phases: 1) altitude mining and 2) proximity mining. These two data mining phases are performed separately for the data from each aircraft type in the ADS-B data set. The phases must occur in sequence to produce correct results. Both of the data mining phases operate on the same relational database. This relational database stores temporary results during each of the data mining phases, as well as permanently storing the best solution to each cluster analysis problem. Data are grouped into tables based on their respective aircraft types.

ADM uses a database-oriented implementation of GKA for altitude mining. Either GKA or a database-oriented implementation of the Expectation-Maximization (EM) algorithm [Dempster, *et al.*, 1977; Plant and Böhm, 2010] may be selected for the proximity mining phase. The implementation of the EM algorithm uses a Gaussian mixture model for the probability distributions.

ADM data mines the data for each aircraft type as follows. In the first phase, the subpaths (i.e. the data points in the clusters) discovered in the data for the aircraft type are automatically clustered into k_1 clusters with respect to the altitude at the start of each subpath. In the second phase, the best solution found during the first phase is data mined further. Each of the k_1 clusters from that solution is data mined with respect to the aircraft's proximity to the nearest private airport at the start of each subpath. This subdivides each of the k_1 altitude clusters into k_2 proximity clusters, resulting in a cluster hierarchy.

4 Results

ADM preprocessed the raw ADS-B data set, which consists of approximately 22 gigabytes of data, to find subpaths in the normalized flight paths of all the aircraft present in the data set. Then ADM mined the preprocessed data, which consist

of subpaths from 42 different types of ADS-B-equipped aircraft. The data mining results for one of the aircraft types (the Gulfstream IV) are shown in table 1. The results for the Gulfstream IV were chosen because the Gulfstream IV is a private jet considered more likely to fly under VFRs in Class E airspace.

ADM first categorized the data points into one of 2 clusters based on the aircraft's altitude. This resulted in a cluster of low altitudes and a cluster of high altitudes. Then, for each of these clusters, ADM categorized the data points within the cluster into one of 3 subclusters based on the aircraft's proximity to the nearest private airport. This resulted in a cluster of aircraft close to private airports, a cluster of aircraft far from private airports, and a cluster of aircraft midway between close to a private airport and far from a private airport. During data mining, ADM evolved 10 candidate solutions for 200 generations for each aircraft type's set of discovered subpaths. This insured that the total within-cluster variation was near its minimum. It also insured that no overlap occurred between any of the resulting altitude or proximity clusters.

The results from data mining the ADS-B data set were fairly consistent. One or two maneuvers occurred much more frequently within a proximity cluster than the rest. However, the results for the Airbus 340-200 were unusual. Since only 13 normalized flight paths were contained in these data, only 2 digital pheromone trails were discovered. Many of the maneuvers were unclassifiable. Consequently, the proximity subclusters associated with one of the altitude clusters did not contain any classifiable subpaths.

Table 1. The relative frequencies of nine types of maneuvers discovered for the Gulfstream IV.

Altitude Range (meters MSL)	Proximity Range (meters)	Aircraft Maneuver	Relative Frequency
655 – 3,078	798 – 4,895	D	73.3%
		C	26.7%
655 – 3,078	7,536 – 10,143	D	100%
655 – 3,078	14,469	D	60%
		C	40%
3,216 – 5,456	1,181 – 5,401	D	28.6%
		G	28.6%
		A	14.3%
		F	14.3%
		H	14.3%
3,216 – 5,456	5,723 – 9,176	D	71.4%
		C	14.3%
		B	4.8%
		E	4.8%
		I	4.8%
3,216 – 5,456	10,057 – 19,344	D	81.8%
		C	9.1%
		E	9.1%

5 Conclusion

Probabilistic models for pilot behavior in Class E airspace were obtained by data mining ADS-B data sets. The

accuracy of these probabilistic models is rather limited due to the small size of the ADS-B data sets we mined. The largest ADS-B data set we mined only contains 3 months of ADS-B data obtained from base stations around the world. However, the airport database we used for proximity calculations should be relatively complete. We will keep it up to date via the Aircraft Owners and Pilots Association (AOPA) website.

We are currently waiting to obtain a much larger NAS Aircraft Situation Display to Industry (ASDI) data set from the U.S. Air Force. This larger ASDI data set, which will contain data similar to the ADS-B data sets we currently have, will consist of several terabytes of data. Data mining this ASDI data set, which will be substantially larger than our current ADS-B data sets, will allow us to obtain more accurate probabilistic models. However, since mining a massive data set is a lengthy process, ADM will be modified to execute its algorithms in parallel on a Beowulf cluster. Another enhancement to ADM will be integrating new ASDI and/or ADS-B data into the database in an incremental fashion. These enhancements will further improve ADM's run-time efficiency in mining massive ADS-B and ASDI data sets.

Acknowledgments

Some of this work was performed with funding from the U.S. Air Force (contract number FA4861-06-C-C006) and the North Dakota EPSCoR program (NSF grant #EPS-0814442).

References

- Roland E. Weibel and R. John Hansman, Jr. Safety considerations for operation of different classes of UAVs in the NAS. In *Proceedings of the Fourth AIAA Aviation Technology, Integration, and Operations Forum*, pages 20–22, Arlington, Virginia, USA, September 2005. AIAA.
- Ronald Marsh, Kirk Ogaard, Micah Kary, John Nordlie, and Chris Theisen. Development of a mobile information display system for UAS operations in North Dakota. *International Journal of Computer Information Systems and Industrial Management Applications*, 3: 435–443, 2011.
- Procedures for handling airspace matters, Federal Aviation Administration, http://www.faa.gov/air_traffic/publications/atpubs/AIR/a1r1801.html, 2010. Last accessed on March 24, 2011.
- Ronald Marsh and Kirk Ogaard. Mining heterogeneous ADS-B data sets for probabilistic models of pilot behavior. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 606–612, Sydney, Australia, December 2010. IEEE.
- K. Krishna and M. Narasimha Murty. Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 29(3): 433–439, June 1999.
- Yi Lu, Shiyong Lu, Farshad Fotouhi, Youping Deng, and Susan J. Brown. Incremental genetic K-means algorithm and its application in gene expression data analysis, *BMC Bioinformatics*, 5(172), October 2004.
- Joel C. McCall, David P. Wipf, Mohan M. Trivedi, and Bhaskar D. Rao. Lane change intent analysis using robust operators and sparse Bayesian learning. *IEEE Transactions on Intelligent Transportation Systems*, 8(3): 431–440, September 2007.
- David Taniar and John Goh. On mining movement patterns from mobile users. *International Journal of Distributed Sensor Networks*, 3(1): 69–86, January 2007.
- H. Van Dyke Parunak, Michael Purcell, and Robert O'Connell. Digital pheromones for autonomous coordination of swarming UAVs. In *Proceedings of the First AIAA Unmanned Aerospace Vehicles, Systems, Technologies, and Operations Conference*, Portsmouth, Virginia, USA, May 2002. AIAA.
- Yasushi Ando, Yoshiaki Fukazawa, Osamu Masutani, Hirotoishi Iwasaki, and Sinich Honiden. Performance of pheromone model for predicting traffic congestion. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 73–80, Hakodate, Japan, May 2006. ACM.
- Hoang-Nam Chu, Arnaud Glad, Olivier Simonin, François Sempé, Alexis Drogoul, and François Charpillet. Swarm approaches for the patrolling problem, information propagation vs. pheromone evaporation. In *Proceedings of the Nineteenth IEEE International Conference on Tools with Artificial Intelligence*, pages 442–449, Patras, Greece, October 2007. IEEE.
- Wolfgang Narzt, Gustav Pomberger, Ursula Wilflingseder, Oliver Seimel, Dieter Kolb, Jan Wieghardt, Horst Hörtnner, and Roland Haring. Self-organization in traffic networks by digital pheromones. In *Proceedings of the IEEE Intelligent Transportation Systems Conference*, pages 490–495, Seattle, Washington, USA, September–October 2007. IEEE.
- Vijay Kalivarapu, Jung-Leng Foo, and Eliot Winer. Improving solution characteristics of particle swarm optimization using digital pheromones. *Journal of Structural and Multidisciplinary Optimization*, 37(4): 415–427, March 2008.
- Thomas George Kannampallil and Wai-Tat Fu. Trail patterns in social tagging systems: Role of tags as digital pheromones. In *Proceedings of the Fifth International Conference on Foundations of Augmented Cognition*, pages 165–174, San Diego, California, USA, July 2009. Springer-Verlag.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society—Series B*, 39(1): 1–38, 1977.
- Claudia Plant and Christian Böhm. Parallel EM-clustering: Fast convergence by asynchronous model updates. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 178–185, Sydney, Australia, December 2010. IEEE.

Discovering and Matching Service Providers among Heterogeneous Agents

Hao Lan Zhang

NIT, Zhejiang University
Ningbo, Zhejiang, China
haolan.zhang@nit.zju.edu.cn

Clement H.C. Leung

Department of Computer Science
Hong Kong Baptist University
clement@comp.hkbu.edu.hk

Xijin Tang

Chinese Academy of Science
Beijing, China
xjtang@amss.ac.cn

Abstract

Multi-agent systems can solve many complex problems through incorporating various task oriented agents. An efficient agent service discovering and matching process will enhance the cooperation and coordination among heterogeneous agents. Agent cooperation can be achieved by matching requesting agents with service-providing agents, and, through such cooperation, multi-agents can solve a variety of complex problems. Improving the efficiency of the agent service matching and discovering process has thus become an important issue in the multi-agent field. The adoption of an appropriate agent-matching mechanism will enhance agent cooperation and communication efficiency within an agent network. In this paper we develop a new agent-matching algorithm, the *Agent Rank* (AR hereafter) algorithm, which ranks service-providing agents based on their previous cooperation with a nominated requesting agent and the semantic relevance to requests. Through the AR algorithm, the chance of discovering an appropriate corresponding agent is improved. The Agent-Rank algorithm overcomes the problems of agent matching in a large agent network through combining general ranking scores with request-based ranking scores; and our empirical results demonstrate the efficiency of the AR algorithm.

1 Introduction

A digital ecosystem may contain diversified types of entities carrying out a variety of functions. Some of these entities are relatively passive (e.g. image annotations and multimedia objects [Wong et al., 2008]) while others are active and serve to cause significant changes to the environment. Of the latter type, agents that deliver useful business services are vital to a productive and sustainable environment. Through the services rendered by such social and autonomous agents, and the resources consumed by them, values are added to different parts of such ecosystems. Multi-agent systems can solve many complex problems by incorporating various task-oriented agents, which represent a problem-solver in each subtask domain. In this way, a complex problem can be

tackled with a high degree of efficiency and flexibility. Thus applying an effective agent cooperation and communication process is the key to a multi-agent system. A multi-agent system must provide some competent agent discovering and matching mechanisms in order to improve the communication and cooperation efficiency amongst agents, and it also should enable agents with a local view of their environment to take action that advance global system objectives [Parunak et al., 2005]. Thus, agents are enabled to perform service discovery efficiently, and the chance of discovering new services are increased.

A number of agent-matching mechanisms have been developed for agents interaction and discovery [Sycara et al., 1999; Sycara and Widoff, 2002; Li et al., 2006]. However, current knowledge-based reasoning mechanisms have insufficient capability to provide an efficient semantic match based on different information resources [Halevy, 2001; Washe, 2001]. This affects the quality and accuracy of an agent matching process. Existing agent matching mechanisms are somehow limited by current semantic reasoning technology. It is commonly admitted that current semantic matching mechanisms are still limited when dealing with complex problems. Therefore, the agent-matching process has to be highly tolerant. The expectation of exact matches in every matching process is not realistic. However, narrowing down the matching objects in a complex matching process can significantly improve matching efficiency. This is one of the major motivations of this paper.

Regular match and re-match processes occur constantly in an agent matching process. Thus, it is more practical to have some mechanisms that could provide the information on related service provider agents (P-agents) and rank these P-agents according to their previous service quality, service frequency, etc. Hence, a request agent (R-agent) can avoid massive searching among various P-agents by selecting a P-agent from the top ranking score. Thus, we propose the new AR algorithm in this paper. The use of the AR algorithm can enhance the efficiency of the existing agent-matching process. The AR algorithm is inspired by the PageRank algorithm [Page et al., 1998].

The system framework for the AR algorithm combines peer-to-peer topologies and the centralised topologies [Zhang et al., 2008; Zhang et al., 2006]. However, the algorithm can be used in a general system framework through

adopting more complex Agent Association Graphs (AAGs). The primary concept of the AR algorithm is to rank all the P-agents in a multi-agent network. According to the ranking process, an R-agent is able to select its P-agent(s) more efficiently and accurately.

The decision-maker (human or agent) is the ultimate selector to choose the most relevant P-agents according to their ranks in the AR algorithm. The default process will select an agent with highest rank score as a P-agent. Thus, the current matching processes will search for P-agents according to the provided ranks instead of massive searching.

2 Related Work

Existing agent-matching methods deploy description language-based matching, such as LARKS (Language for Advertisement and Request for Knowledge Sharing) [Sycara et al., 1999; Sycara and Widoff, 2002], SDL (Service Description Language) [Subrahmanian et al, 2000], etc. Most of these methods rely on matching agent-service-descriptions, also called service advertisements. These descriptions indicate an agent's capabilities and services. An R-agent can search for a P-agent through matching the descriptions provided by the other agents. In LARKS, there are three matching mechanisms including exact match, plug-in match, and relax match. Exact match is the most accurate matching, which requires that both service descriptions being literally equal. Plug-In match is a less accurate matching method in LARKS that two matching descriptions are generally similar. Exact match is a special case of plug-in match. Relax match is the least accurate matching method. It determines the similarity of two descriptions by returning a numerical distance value instead of semantically matching the two descriptions. Semantic matching is essential to relax match.

The above three matching mechanisms are essentially based on semantic matching, which involves searching for an appropriate agent from their information providers. These mechanisms often neglect the uncertainty factors in the matching process. For instance, a plug-in match process might match an appropriate agent from a large multi-agent system. In fact, thousands of agents can fulfil the search requirement and the first matched agent might not be the most appropriate P-agent. Unfortunately, most existing agent-matching mechanisms do not consider this circumstance and this could increase the failures in the matching processes. Moreover, current agent-matching mechanisms often neglect some important factors, such as response quality, visiting history, etc. The use of the AR algorithm could eliminate these problems by ranking all the potential P-agent candidates and considering various factors that affect agent matching efficiency. The AR algorithm is a method to improve the efficiency of existing agent-matching mechanisms. In particular, the re-matching process is highly efficient in AR-based systems. The re-match process in the AR algorithm does not require searching all the agent information providers again. It simply continues its operation based on the ranking list generated previously.

PageRank algorithm is the pioneer work of the AR algorithm. However, they are very different from each other. (i)

Semantic matching is the core in PageRank process, whereas the AR algorithm combines both semantic-based matching and request-independent ranking processes. (ii) PageRank algorithm described in [Page et al., 1998] does not consider any response quality issue in its matching process. On the contrary, the response quality factor is one of the major factors that affect an agent's ranking score in the AR algorithm. (iii) PageRank forms a probability distribution over Web pages. Different from PageRank, the AR algorithm indicates the ranks of the P-agent candidates associated with an R-agent. The frame of reference in PageRank is an overall network, whereas it is a specific R-agent in the AR algorithm.

Previous research work has suggested several ranking mechanisms based on heterogeneous agents such as [Kuhn, 1993]. However, the comparative ranking function mainly deals with goal-oriented tasks and neglects some important factors in agent cooperation such as response quality, visit frequency, etc. In addition, the early studies do not distinguish P-agents and R-agents, which complicates the ranking process and increases ranking overheads since ranking objects involve all agents in a system instead of a group of related P-agent candidates. This results in the comparative ranking process becoming complicated when a large number of agents are involved.

3 General ranking in AR Algorithm

In an AR-based multi-agent system, the agent-ranking process consists of two major stages. The first stage is called the General ranking stage, which ranks the importance of all the agents that have relationships with the R-agent.

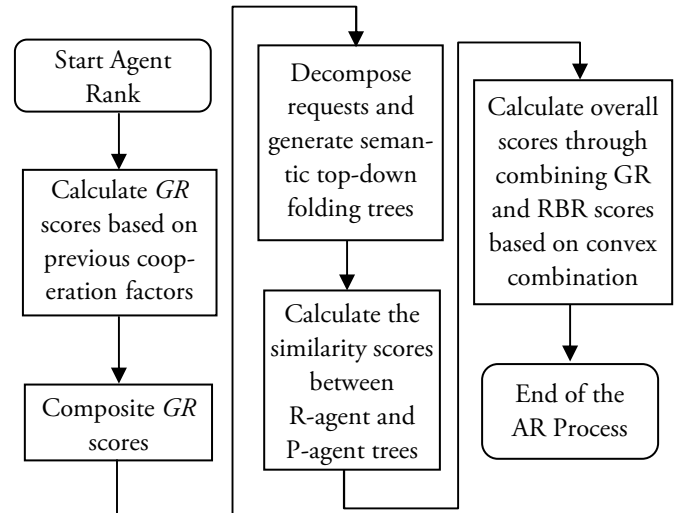


Figure 1. Major procedures of the AR-based Process

The second stage of the AR process is Request-based ranking, which further focuses on specific searching and matching based on a request and incorporates other major factors such as visiting history, quality of service and the results from the general ranking stage. Figure 1 illustrates the major steps in the Agent Rank process.

The general ranking process ranks the importance of all the agents related to the R-agents. The related agents refer to

those agents that send or receive requests from or to the request agent. The general ranking process adopts the basic idea of PageRank that calculates agents' incoming and outgoing links to form an importance distribution over an agent network. Unlike PageRank, the visit frequency of an R-agent to a P-agent is calculated in the general ranking process because the frequency of usage of a P-agent is an important criterion for evaluating an agent's importance. An AAG is used to analyse the ranking process. The AAG represents the cooperation history of an R-agent with its associated agents.

The first step in analysing the general ranking process is generating an AAG. Normally, an R-agent and its associated agents (P-agent candidates) form a Small world topology graph [Watts and Strogatz, 1998; Walsh, 1999]. The links associate the P-agents and R-agent. Incoming links in an AAG denotes an agent served as a P-agent in the previous cooperation with the origin agent of the link. The outgoing link denotes an agent that served as an R-agent in the previous cooperation with the destination agent.

Most of the network graphs in the general ranking process can be transformed to a small-world graph. In such a graph, the R-agent is always the centre of the graph. The links in these graphs represent previous cooperation between the R-agent and the other agents. The general ranking process ranks all the P-agent candidates in the R-agent-based graph. The ranking process is based on the calculation of the total inbound links of the P-agents.

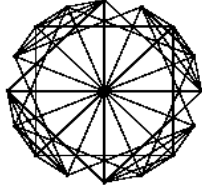


Figure 2. A small-world form of AAG [Walsh, 1999].

3.1 The ranking calculation in General Ranking

The detailed factors that need to be considered in this ranking process including the total number of links in a graph, the damping factor, the visit frequency by the R-agent, the total visit numbers in a period, and the response quality. The total number of links, which can be systematically evaluated node by node, is:

$$E = \frac{\sum_{i=1}^m (G_i + C_i)}{2} \quad (1)$$

where E denotes the total number of links in an AAG; G_i denotes the outbound links of an agent; C_i denotes the inbound links of an agent in an AAG; and m denotes the total agent number, including the R-agent, in an AAG. The reason for Equation.1 is that: the inbound links of each node (agent) are added to the outbound links to generate the total number of links $\sum_{i=1}^m (G_i + C_i)$.

When the total link number of all the agents is added, the number is duplicated because one link has two nodes and

thus each link is counted twice. Therefore, the number needs to be divided by 2, which is the total number of links.

The general ranking process adopts the damping factor, which is different from its description in the *PageRank* algorithm [Page et al., 1998]. The damping factor in the general ranking process is based on the ratio of the total number of an agent's inbound links to the total number of links of all agents associated with the R-agent (including the R-agent) and details can be found in [Zhang et al., 2010].

The visit frequency (V_F) coefficient of a P-agent candidate provides information about how often the nominated R-agent has associated with this P-agent candidate previously. It is a request-independent factor, which is not affected by the requests sent by an R-agent. V_F helps the AR process determine whether an agent's capability is out of date. There are two principles we employ in this stage including: (i) only the P-agent candidates' inbound links from the R-agent are considered, (ii) only recent visits are considered (subject to specific organisational cases). The coefficient of a P-agent candidate's V_F is calculated as follows:

$$f(V_p) = 1 + (V_p / \sum_{i=1}^n V_i) \quad (2)$$

where $f(V_p)$ denotes the coefficient of recent visits from an R-agent to a nominated P-agent candidate; V_p denotes the total number of recent visits from an R-agent to its nominated P-agent candidate p ; n denotes the total number of P-agent candidates that have inbound links from their R-agent; V_i denotes the total number of recent visits from an R-agent to all its P-agent candidates on an AAG ($i = 1, 2, 3, \dots, n$).

Equation 2 can be further explained that the coefficient of the V_F is important to the general ranking score; however, its impact on the score should not be unlimited. Therefore, the adoption of the ratio of V_p to V_i limits the impact of the V_F factor and the ratio is always between 0 and 1. In addition, the V_F factor is positive such that if there are more visits from the R-agent to its P-agent candidate, then this P-agent's general ranking score should be higher. The coefficient is always greater than 1. For a P-agent candidate that only has outbound links to its R-agent, then $f(V_p) = 1$.

The response quality factor provides information about the quality of the service provided by a P-agent to its R-agent previously. It can be expressed as follows:

$$Q_p = \left(1 - \frac{(T_p - T_{avg})}{T_p} \right) \times \left(\left(1 + \frac{1}{L_p} \right) / 2 \right) \\ = (T_{avg} / T_p) \times (0.5 + 1 / 2 L_p) \quad (3)$$

where Q_p denotes P-agent candidate X's average response quality to an R-agent; T_i denotes each individual P-agent candidate's average response time to its R-agent and i is a variable; T_p denotes X's average response time to its R-agent; L_p denotes the number of links that X established with other agents to solve a request from its R-agent.

The response quality issue in the agent matching area has been raised in previous work such as [Sycara and Widoff, 2002].

3.2 General Ranking Composition

Based on the previous equations, the general ranking equation can be deduced as follows:

$$GR(A) = \underbrace{\left((1-D) + D \times \left(\sum_j \frac{GR_j}{G_j} + GR_{RA} \right) \right)}_{\text{Part 1}} \times \underbrace{Q_A \times f(V_A)}_{\text{Part 2}}$$

where $GR(A)$ denotes the general ranking score of a P-agent candidate A (A hereafter); GR_j denotes the general ranking score of P-agent candidate j that points its outbound link towards A (votes A); G_j denotes the outbound links of P-agent candidate j ; j is a variable that denotes the P-agent candidates vote A ; GR_{RA} denotes the ranking score of the R-agent that has outbound links to the P-agent candidates; Q_A is the response quality of P-agent A to its R-agent; $f(V_A)$ is the coefficient of A 's visit frequency. If the selected P-agent has no inbound links from its R-agent, then $GR_{RA} = 0$. Otherwise,

$$GR_{RA} = \sum_j \frac{GR_j}{n}$$

where n denotes the total number of P-agent candidates; GR_{RA} denotes the general ranking score of an R-agent. R-agent's general ranking score is the average value of all its P-agent candidates' general ranking score. The reason to count R-agent's rank into its general ranking score is: if a P-agent candidate receives inbound links from its R-agent then the general ranking score of this agent increases. Thus, the R-agent has an impact on a P-agent's general ranking score. Note that the determination of $GR(A)$ is an iterative process, since GR_j in the equation will in turn depend on $GR(A)$.

The above general ranking equation can be further expanded as follows. This further incorporates the Q and $f(V)$ factors in detail.

$$GR(A) = \left(1 + \frac{2C_r}{\sum_{i=1}^m (G_i + C_i)} \times \left(\sum_j \frac{GR_j}{G_j} + GR_{RA} - 1 \right) \right) \times \frac{T_{avg} \times \left(1 + \frac{1}{L_p} \right)}{2T_p} \times \left(1 + \frac{V_p}{\sum_{i=1}^n V_i} \right)$$

4 Request-based Ranking

The request-based ranking process consists of three major phases including generating agent capability semantic tree, generating request-based semantic tree, and Request-Based Ranking (RBR) similarity calculation. The first two steps are quite similar and they can be illustrated in Figure 3. The process of constructing a RBR semantic extension tree or an agent capability description tree is based on existing semantic networks [Bhavsar et al., 2004] and ontology

technology. There are several reasons for adopting a tree-based similarity calculation method in agent matching in preference to using existing text mining or Web-based semantic matching technology; and they are listed as follows:

- Semantic-tree description is a flexible and extensible mechanism for semantic phase decomposition.
- Existing agent description languages cannot distinguish agent capabilities efficiently as many agents, especially in a large multi-agent system, seem to provide very similar descriptions. Therefore, an agent matching process requires more accurate and detailed analysis compared to simple text mining.
- The semantic-tree can generate independent keywords that are efficient for similarity calculation.

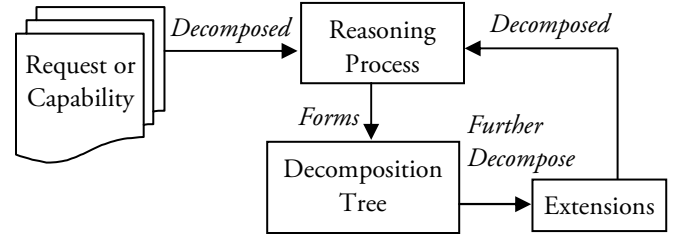


Figure 3. Generating a semantic extension tree

4.1 RBR Calculation Process

The RBR calculation process deploys an extraction mechanism to collect two information sets from two trees before calculating the similarity between a RBR semantic extension tree and an agent capability description tree. Each set contains a number of semantic elements and coefficients representing the node information in both trees. The similarity calculation is based on semantic matching between the two sets and coefficient calculation of the matched elements. The RBR similarity calculation process makes use of some concepts from existing data mining and information retrieval technologies including keyword-based retrieval, similarity-based retrieval and document classification analysis [Han and Kamber, 2001].

The Cosine-Similarity Measure (CSM) has been applied extensively in the information retrieval domain to calculate the similarity between two vectors [Ganesan et al., 2003], which is their inner product normalised by their lengths,

$$sim_{Cos}(X, Y) = \frac{\vec{X} \cdot \vec{Y}}{|\vec{X}| |\vec{Y}|} \quad (4)$$

However, the CSM cannot function efficiently when the dimensions of these vectors are unequal. To overcome this, we adopt the Cartesian-Product [Hein, 2002] method in the RBR-based similarity calculation between two sets, which is defined as:

$$A \times B = \{ (a, b) | a \in A \text{ and } b \in B \} \quad (5)$$

where A contains the elements extracted from a RBR semantic tree and B contains the elements extracted from an agent capability description tree.

Due to the space limitation, this paper illustrates the concept of the RBR-based similarity calculation method instead of providing the detailed calculation procedures. Detailed information can be found in [Zhang et al., 2010]. The RBR-based similarity calculation consists of three main steps. The first step generates a Cartesian product; the second step calculates the similarity value of two elements in the Cartesian product; the final step sums the level coefficient and the path coefficient of a pair of matched elements through incorporating their similarity values. For instance, consider two sets $S1$ and $S2$, which are extracted from a RBR semantic extension tree ($S1$) and an agent capability description tree ($S2$), respectively.

In brief, the RBR process ranks the similarity scores of all the agent capability description trees. The final agent ranking scores combine the general ranking scores and the RBR similarity scores.

4.2 RBR Calculation and Experimental Results

The following example illustrates the RBR-based similarity score calculation. Figure 4 shows an RBR-based semantic extension tree, which is based on a request from the R-agent (R-A). Figure 5 shows an example agent capability description trees for a specific agent.

Requests for Y Institute Salary Statistics in 2010

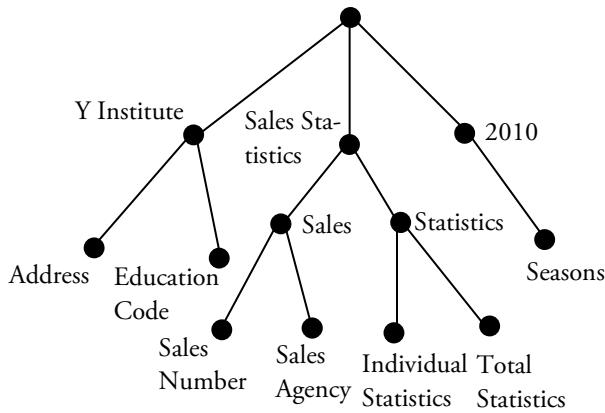


Figure 4. An RBR semantic extension tree based on R-A

Y Institute Salary Management

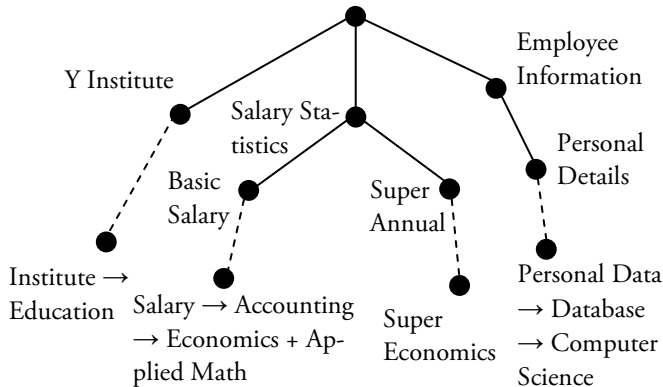


Figure 5. An example agent capability description tree

The authors have developed an experimental system based on the request-based ranking methodology. The system incorporates the similarity relationships based on *WordNet* [Li et al., 2003]. The experiment is based on 4 P-agents and 1 R-agent. The RBR-based similarity scores of the example in the different levels have been generated and shown in Figure 6.

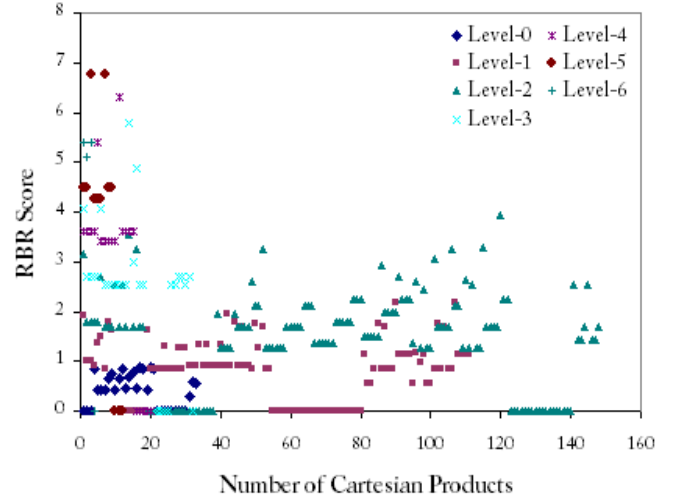


Figure 6. RBR similarity scores of a selected P-agent: A1

The node numbers at different tree levels of the above examples are different and this affects the RBR similarity scores at different levels as shown in Figure 6. There are more nodes in levels one and two than in the other levels in the above four P-agent candidate trees. The general scores in Figure 6 are higher than those in other agents. In other words, agents A1 offer more relevant functions to solve the request from request agent R-A than the other agents. The average scores of the above agents to the R-agent are listed below (in descending order of scores):

Agent Name	RBR similarity score
A1	1.38
A3	1.10
A2	1.06
A4	1.01

Thus, A1 is the most relevant agent in dealing with the request from R-agent.

5 Discussion

The Agent-Rank algorithm consists of the general ranking score and the request-based ranking score. The proportion of each ranking score in the overall ranking score depends on specific applications. In general, the overall agent-ranking score can be calculated using the convex combination:

$$AR = \alpha \times GR + (1 - \alpha) \times RS \quad (6)$$

where α ($0 < \alpha < 1$) is the weight assigned to GR , which denotes the general ranking score; RS denotes request-based ranking score which receives a weight of $(1-\alpha)$; and the AR gives the overall agent-ranking score. Supposing the exam-

ples of the previous section are based on a large multi-agent system, then the general ranking score should be assigned more weight (i.e. $\alpha > 0.5$). Using $\alpha = 65\%$, which is the proportion based on the calculation of the total number of agents and average request number, the scores in the example of the previous section can be summarised as:

$$A1 = 5.73 \times 0.65 + (1-0.65) \times 1.38 \approx 4.21,$$

$$A2 = 1.45 \times 0.65 + (1-0.65) \times 1.06 \approx 1.32,$$

$$A3 = 5.80 \times 0.65 + (1-0.65) \times 1.10 \approx 4.16,$$

$$A4 = 0.56 \times 0.65 + (1-0.65) \times 1.01 \approx 0.72.$$

Thus A1 is the most appropriate P-agent to R-agent RA among the four trees. Thus, A1 is selected as a P-agent.

6 Conclusion

The agent-matching problem has a different level of complexity compared with the searching and matching of web pages. Over and above the linear matching of strings to determine relevance, the discovery of functions and capabilities is required. Here, by adopting a collaborative mission oriented focus, we use a methodology whereby past relationships and performance are taken into account to enhance the ability to achieve mission objectives.

The AR algorithm provides an efficient methodology to discover an R-agent's request within numerous P-agents. It adopts and modifies the concept of the PageRank algorithm. The AR algorithm further improves matching through taking the response quality issue into account in its matching process. The application of AR algorithm in agent service discovery processes can efficiently improve the overall system performance, which has been proved by the experiments. The Agent-Rank algorithm avoids massive search among numerous agents, and is performance efficient in a large multi-agent system. Although the present algorithm does carry some ranking computation overhead, such computation cost is considered acceptable in comparison with the penalty of inappropriate matching which may result in defective solutions and time wasted in carrying out re-matching.

References

- [Wong et al., 2008] Wong, R. C. F. and Leung, C. H. C. "Automatic semantic annotation of real-world Web images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 11, 2008, pp. 1933-1944..
- [Parunak et al., 2005] Parunak, H.V.D., Brueckner, S.A., and Matthews, R. "Pheromone Learning for Self-Organizing Agents", *IEEE Transactions on SMC – Part A*, Vol.35, No.3, 2005, pp. 316-326, IEEE CS Press.
- [Sycara et al., 1999] Sycara, K., Klusch, M., and Widoff, S. "Dynamic Service Matchmaking Among Agents in Open Information Environments," *SIGMOD Record*, Vol. 28, No.1, 1999, pp.47-53.
- [Sycara and Widoff, 2002] Sycara, K., and Widoff, S. "LARKS: Dynamic Matchmaking Among Heterogeneous Software Agents in Cyberspace," *Autonomous Agents and Multi-agent Systems*, 5, 2002, pp. 173-203. Kluwer Academic Publishers.
- [Li et al., 2006] Li, X., Montazemi, A. R., and Yuan, Y. "Agent-based buddy-finding methodology for knowledge sharing," *Information & Management*, Vol. 43, 2006, pp.283-296. Elsevier Press.
- [Halevy, 2001] Halevy, A. "Answering queries using views: a survey," *VLDB Journal*, 10 (4), 2001, pp. 270 – 294.
- [Washe, 2001] Washe, H. et al., "Ontology-based integration of information-a survey of existing approaches," *Proc. of IJCAI*, August 2001.
- [Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. "The PageRank citation ranking: Bringing order to the web," *Technical report*, Stanford University, Stanford, CA. 1998.
- [Zhang et al., 2008] Zhang, H.L., Leung, C.H.C., and Raikundalia, G.K. "Topological analysis of AOCD-based agent networks and experimental results," *Journal of Computer and System Sciences*, Vol. 74, 2008, pp. 255 – 278, Elsevier Publication.
- [Zhang et al., 2006] Zhang, H.L., Leung, C.H.C., and Raikundalia, G.K. "Matrix-Agent Framework: A Virtual Platform for Multi-agents", *Journal of System Sciences and Systems Engineering*, Vol. 15, No. 4, 2006, pp. 436 – 456, Springer Press.
- [Subrahmanian et al, 2000] Subrahmanian, V., Bonatti, P., Dix, J. et al. *Heterogeneous Agent Systems*. 2000, pp.43-59. MIT Press.
- [Kuhn, 1993] Kuhn, N. "Comparing Ranking of Heterogeneous Agents," *Proc. of COOCS*, 1993, pp. 1-12. ACM Press.
- [Watts and Strogatz, 1998] Watts, D.J. and Strogatz, S.H. "Collective Dynamics of 'Small world' Networks," *Nature*, Vol.393, 1998, pp. 440-442.
- [Walsh, 1999] Walsh, T. "Search in a Small World," *Proc. of IJCAI*, 1999, pp.1172-1177. Morgan Kaufmann Press.
- [Zhang et al., 2010] Zhang, H.L., Leung, C.H.C., Raikundalia, G.K. and He, J. "A Novel Ranking Algorithm for Service Matching Based on Agent Association Graphs", *Workshop Proc. of IEEE ICDM*, 2010, pp. 1273 – 1280, Sydney, Australia, IEEE Press.
- [Bhavsar et al., 2004] Bhavsar, V.C., Boley, H., and Yang, L. "A Weighted-Tree Similarity Algorithm for Multi-agent Systems in E-business Environments," *Computational Intelligence*, Vol. 20, No. 4, 2004, pp. 584-602.
- [Han and Kamber, 2001] Han, J. and Kamber, M. *Data Mining: Concept and Techniques*. 2001, pp. 428-440. Morgan Kaufman Publishers.
- [Ganesan et al., 2003] Ganesan, P., Garcia-Molina, H., and Widom, J. "Exploiting Hierarchical Domain Structure to Compute Similarity," *ACM Transactions on Information Systems*, Vol. 21, No. 1, 2003, pp. 64-93. ACM Press.
- [Hein, 2002] Hein, J.L. *Discrete Mathematics*. 2002, pp. 74-75, 140. Jones and Bartlett Publishers.
- [Li et al., 2003] Li, Y., Bandar, Z.A., and McLean, D. "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transaction on Knowledge and Data Engineering*, Vol.15, No.4, 2003, pp. 871-882. IEEE CS Press.

Extraction of Useful Utterances from Discussion Text for Improving Value of Idea

Yoko Nishihara and Yukio Ohsawa

Department of Systems Innovation, School of Engineering, the University of Tokyo

7-3-1, Hongo, Bunkyo, Tokyo 113-8656, Japan

nishihara@sys.t.u-tokyo.ac.jp, ohsawa@sys.t.u-tokyo.ac.jp

Abstract

Creative activity performed by a group of people needs human communications in which people discuss and think up new ideas for solving their problems. The communication is useful for finding defects and improvements of new idea. Many business companies tend to record communications of idea creation and transcribe the recorded communications. The transcribed texts contain useful utterances for idea creation and improvement. However, the transcribed texts are merely read because of the low amount of contained information. New methods for extracting useful utterances from transcribed texts are required. This paper proposes a new method for extracting useful utterances from a communication text. We verified that extracted utterances by our method were more useful than utterances extracted by the previous method of text summarization.

1 Introduction

Creative activity is a process of making new and important products and items [6]. We see many creative activities in business companies making new commercial items and new industrial products. Research activities in academic universities and colleges are also the creative activities.

People doing creative activity have purposes to be achieved. People think up new ideas to achieve their purposes, then create new products and items based on thought up ideas. The process of creative activity is run by running over the two actions. If people obtain valuable ideas, they can run the process smoothly. The most important thing is to obtain valuable ideas.

When thinking up a new idea, a group discussion often works well. The proverb *four eyes see more than two* represents the goodness. A group discussion gives hints to people which are obtained from individuals' viewpoints. The more hints are obtained, the more ideas are improved into new ideas with less defects. Opinion exchanges in a discussion are made naturally by people for obtaining good ideas.

In a group discussion, useful information is provided to rise the value of idea. In this study, the value of idea is defined as originality and feasibility of idea which has

been adopted by previous researches in creative activity[4; 18]. Business companies have aggressively recorded transcribed texts of discussion to utilize in the future discussion [16]. The transcribed texts, unfortunately, will rarely be utilized. Because information obtained from reading the transcribed texts is fewer though it takes much time to read the texts. It is necessary for rising up the value of idea to utilize the texts in the future discussion. It is also necessary for utilizing the texts to organize the texts by extracting useful utterances from the discussion texts.

This paper proposes a new method for extracting utterances including useful information from conversation records. We focus on conversations in which new ideas for creative activity are thought up. The authors think that ideas are not valuable just after being thought up. It is necessary to be reconsidered by referring other's comments for obtaining valuable ideas. For example, good research themes can not be obtained without any discussions by members of research team. When a researcher thinks up new a new research theme, he/she have to discuss the new research theme to be improved by other's comments and criticisms. If he/she gets many comments and criticisms, the comments and criticisms should be ordered according to importance to be solved. The motivation of our research is to improve thought up ideas by extracting useful information from conversational records.

2 Related Work

Our work is related to the previous works about structuration and visualization of discussion, communication in creative activity, and text summarization. This section introduces the previous works and shows the differences between the previous works and our work.

2.1 Structuration and Visualization of Discussion

Discussions have to be held over and over again. Reflecting the transcribed texts of past discussions is necessary to make future discussions fruitful ones.

To support for reflecting past discussions, the simplest way is to write and publish transcribed texts of past discussion. However, the amount of words included in a transcribed text is too much to read. The important utterance which should be reflected is few in a transcribed text. The texts have to be organized again.

Many researchers have proposed methods of structuration and visualization of discussion [11; 12; 14]. The previous methods can automatically extract related utterances from a text to represent the relationships among utterances in a discussion text. The methods can also detect separating points of discussion text. These functions of structuration and visualization assist people in understanding discussion flows and discussion perspectives. Then, people can revolve new discussions based on findings and knowledge obtained from reflections of the past discussions.

On the other hand, the previous works have been focused on understanding discussion perspectives, not on understanding discussion details such as important utterances. When supposing discussions for thinking up new ideas for creative activity, not only discussion perspectives, but also discussion details such as what opinions have been given to ideas are quite important. If people can grasp detailed information easily, reflections of past discussion are more supported, and idea thinking is also supported. We propose a new method for extracting utterances including information which will be useful for rising up the value of new idea. The method matches extracted utterances with the thought up ideas for organizing structured discussion texts.

2.2 Communication in Creative Activity

As one of the communication methods for thinking up new ideas, brainstorming has been used for a long time [17]. Brainstorming is a group creativity technique designed to generate a large number of ideas for the solution of a problem. There are four basic rules in brainstorming. These are intended to reduce social inhibitions among group members, stimulate idea generation, and increase overall creativity of the group: (1) withhold criticism, (2) welcome unusual ideas, (3) focus on quantity, and (4) combine and improve ideas. Brainstorming emphasis on the quantity of ideas as shown by the rule (3). Therefore, the rule (1) and the rule (2) for making no limitation in people's thoughts have been set. Even if people obtain many ideas, only a few ideas can be utilized in creative activity. Moreover, people have to consider the quality of idea for achieving the purpose of creative activity. To obtain valuable ideas utilized in creative activity, it is considered that people need to make new discussions without the rule (1) and the rule (2).

What can people do to rise up the value of idea? The theory advocated by Popper who was a researcher in the area of scientific philosophy should be referred. Popper said that the value of theory was risen by obtaining criticisms and defeating the criticisms [20]. The theory by Popper has been limited to the value of scientific theory. It is considered that his theory can be applied to the value of idea in creative activity. The value of idea thought up in creative activity may be risen up by obtaining and defeating the criticisms.

Criticisms are sometimes given by asking questions. For example, a question *your system can be used in this situation?* is considered as a criticism *your system is not considered to be used in this situation*. If the criticisms and questions are used effectively in discussions thinking up new ideas, it has been discovered that the value of idea is risen [15]. Eris has been discovered that there were two questions which were good

for industrial design: Deep Reasoning Question and Generative Design Question [3]. When obtaining requirements from customers, positive/negative questions to intention, purpose, and technique to apply requirements are possibly able to reveal constraints having by customers [9]. These researches intended to follow Popper's research.

It has been discovered that utterance styles of criticism and question are useful for rising up the value of idea thought up by people. However, few researches have tried to find linguistic features of utterance style. Therefore, this paper finds the linguistic features and proposes a new method for extracting effective utterances of criticism and question. We will accelerate to understand the details of discussion by structuring, and to make reflections of discussion.

2.3 Text Summarization

Automatic text summarization has been achieved by extracting important sentences from a text. The importance of sentence is evaluated by words included in a sentence. If a sentence includes important words, a sentence is judged as a important sentence. The importance of word, for example, is evaluated by its frequency in a text [10] and its inverse document frequency among documents [22].

Not only word frequency, but also semantic information has been also utilized for making a summary of text [19]. The semantic informations are, for example, a lexical chain of words [1], stochastic graph of words [2], latent semantics [7], and so on. These methods have succeeded to rise up the accuracy of information retrieval on the Web. The automatic and intelligent methods enables us to obtain appropriate information immediately.

The previous methods have been originally developed for summarizing a text, not multiple texts. Nowadays, the methods and improved methods are being utilized for making summaries of multiple texts [5; 13; 21]. People can manage a huge amount of texts as a compact information by using the summarization methods.

It is easy for summarization to extract important sentences from a text. When people produce summaries, they do not simply extract sentences and concatenate them. In making summaries of multiple documents, automatic sentence comprehension also makes a significant contribution [8]. Our goal is also to extract important information from a text. On that point, our method can be regarded as a summarization method of text. However, our method extracts utterances which may rise up the value of idea though the previous summarization methods have extracted sentences which may be useful for understanding a text. The extracted sentences by the previous method and the extracted utterances by our method are different from each other on the point of their utilization.

3 Discussion Model: Bitters Do Good to the Stomach

Figure 1 shows our discussion model. The model consists of four factors: a discussion topic, sub-topics relating a discussion topic, attendees to a discussion, and styles of utterances

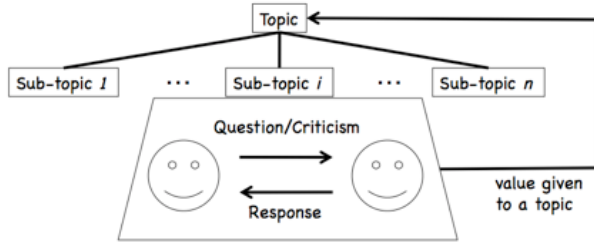


Figure 1: Discussion model: Bitters Do Good to the Stomach.

given to a discussion. The attendees give sub-topics respectively and discuss each other. All sub-topics are related to a discussion topic. However, all sub-topics do not always rise the value of topic.

We suppose that when attendees discuss, an utterance which style is criticism and question is uttered to a sub-topic. It is considered that when criticisms and questions are given, that indicates that attendees find defects and insufficient conditions to a sub-topic. That means given criticisms and questions support attendees for knowing unknown conditions. The unknown conditions can be considered as possibilities for rising up the value of topic. Moreover, if a discussion relating to a criticism and a question ends shortly, a criticism and a question do not give new viewpoints to attendees. On the other hand, if a solution to a criticism and a question is not given immediately and a discussion relating to a criticism and a question continues for a long time, a criticism and a question may give new viewpoints to attendees, i.e., the value of topic may be risen.

We name the discussion model *Bitters Do Good to the Stomach*. The reason is that strict criticisms and questions which can not be agreed easily exactly contribute to rise up the value of idea.

4 Extraction Method of Useful Utterance from Discussion Text

This section explains our extraction method of useful utterances from a discussion text. The method is developed based on the discussion model introduced in Section 3. The outline of the proposed method is shown in Figure 2. A discussion text in which new ideas are thought up is input to the proposed method. The method extracts utterances which have effects for rising up the value of idea from a discussion text. Then, the method output the extracted utterances.

4.1 Algorithm of Extracting Utterances from a Text

This section shows our algorithm of extracting utterances from a text. The flow chart is shown in Figure 3.

step 1 Input a text T

step 2 Separate a input text into small texts $t_i (i = 1, \dots, M)$. A small text has a sub-topic s_i including utterances $u_{ij} (j = 1, \dots, N)$.

step 3 If $i \leq M$, then check whether or not each utterance style is a criticism and a question. If all utterance styles

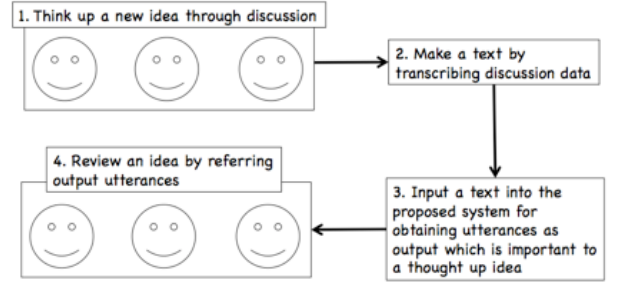


Figure 2: Outline of proposed method.

are not criticisms or questions, then go to step 3. Otherwise, go to step 4.

step 4 If the number of utterances which styles are criticisms and questions is less than a threshold K_1 , increment i to $i+1$, then go to step 3. If the number is more than K_1 , extracts utterances whose length are more than a threshold K_2 . Increment i to $i+1$, then go to step 3.

The threshold K_1 is used for judging whether or not solutions to criticisms are given immediately. If the number of utterances relating to a sub-topic is less than the threshold K_1 , the algorithm can judge that a solution is not given immediately. The threshold K_2 is used for judging which utterances are the most concrete utterances which styles are criticisms and questions. If a utterance does not point out the defects of sub-topic concretely, a utterance does not have an efficiency for rising up the value of topic. The threshold K_2 can extracts utterances for rising up the value of topic.

5 Experiment

We experimented with the proposed method. The purpose of experiment is to verify that the proposed method can extract utterances which are useful for rising up the value of idea.

5.1 Experimental Procedure

The procedure of experiment was as follows:

exp 1 Transcribed a recorded discussion about idea thinking up into a text.

exp 2 Extracted utterances from a text which are useful for rising up the value of idea.

exp 3 Compared extracted utterances with utterances which were judged as useful by participants.

Table1 shows texts used in the experiment. When inputting a text into the proposed method, an idea was regarded as a topic. We made small texts including utterances relating to a idea, and then input small texts into the proposed method.

To evaluate the efficiency of extraction, we prepared a baseline method which extracts utterances by evaluating word weights by tfidf [22]. The baseline method extracted equal number of utterances of the proposed method.

To compare the efficiencies between the proposed method and the baseline method, we calculated precisions and recalls of utterance extraction. The precision and the recall were calculated by using Equation (1) and Equation (2).

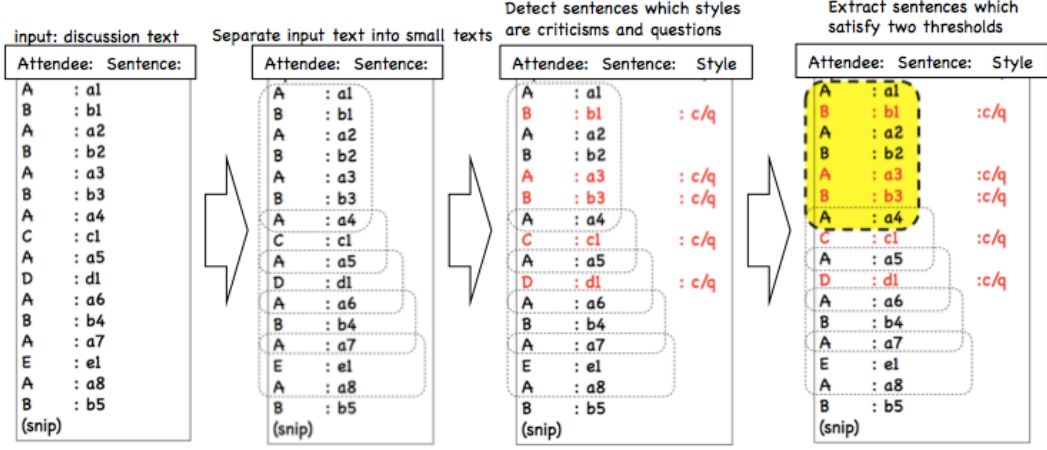


Figure 3: Flow chart of our extraction algorithm of utterances from a discussion text.

Table 1: Text used in experiment. The number of attendees, the number of ideas, the number of utterances and the number of correct utterances judged by participants are shown.

Text	# of attendees	# of ideas	# of utterances	# of correct utterances
#1	6	16	596	52
#2	7	20	1212	70
#3	12	13	380	23
#4	11	12	321	35
#5	11	17	730	67

Table 2: Precisions obtained from experimental results.

Text	Proposed method			Baseline method		
	One	Two	Three	One	Two	Three
#1	0.70	0.56	0.29	0.41	0.22	0.13
#2	0.73	0.40	0.13	0.22	0.02	0.00
#3	0.74	0.41	0.12	0.35	0.15	0.00
#4	0.95	0.62	0.19	0.33	0.14	0.05
#5	0.80	0.55	0.21	0.45	0.13	0.05

$$precision = \frac{\# \text{ of both}}{\# \text{ of extracted}} \quad (1)$$

$$redall = \frac{\# \text{ of both}}{\# \text{ of correct}} \quad (2)$$

The correct utterances were defined by questionnaires to participants. The number of participants was 25 who were graduate/undergraduate students majoring engineering. We asked the participants to read texts and choose utterances which were judged as useful for rising up the value of idea. We defined three types of correct utterances which were chosen by more than one, two, and three participants.

5.2 Experimental Result

Table2 and Table3 shows experimental results about the precisions and the recalls. The precisions of the proposed method were higher than those of the baseline method ($t = 8.1, p = 0.0012 < 0.01$). The recalls of the proposed method were higher than those of the baseline method ($t = 10.1, p = 0.0053 < 0.01$).

6 Discussion

This section discusses the reasons of obtaining high precisions and high recalls from referring the results of the proposed method and the results of the baseline method.

Table 3: Recalls obtained from experimental results.

Text	Proposed method			Baseline method		
	One	Two	Three	One	Two	Three
#1	0.43	0.67	0.75	0.25	0.27	0.33
#2	0.43	0.56	0.62	0.13	0.03	0.00
#3	0.36	0.61	0.67	0.17	0.22	0.00
#4	0.23	0.37	0.44	0.08	0.09	0.11
#5	0.44	0.66	0.77	0.25	0.15	0.18

6.1 Output by the proposed method and the reason of high accuracy

Example of extracted utterance by the proposed method

Table 4 shows an example of a separated text from #2 text. The sub topic in Table 4 is about idea of new machine putting bath agent into bathtub automatically to assist people's comfortable bath time. The first utterance *I would like to put the bath salt in by myself and not automatically* in Table 4 was a correct utterance. This utterance criticizes that putting bath salt into bathtub by himself/herself is the very enjoyable action. After the utterance, three utterances were uttered, but the problem was not solved. An attendee uttering this utterance might think if this point was improved, the value of idea was risen up. However, it is difficult to improve this point. It is considered that the difficulty is shown in the length of exchanges of utterances.

The result indicates that utterances contributing to rise up the value of idea are usually uttered by criticisms and questions. If it is difficult to solve problems pointed by questions and criticisms, the length of utterance exchanges becomes long. This phenomenon matches with our discussion model. Therefore, the proposed method could extract correct utter-

Table 4: Example of a text from #2 text. A correct and extracted utterance by proposed method is shown in *Italic font*.

Attendee	utterance
A	<i>I would like to put the bath salt in by myself and not automatically.</i>
B	This system can take a solid bath salt, not a liquid one.
A	I would like to watch the broadening of the bath salt by adding myself to a bath.
B	You would like to watch it, wouldn't you?

Table 5: Features and numbers of incorrect utterances extracted by proposed. C&Q means a criticism and a question.

Text	Extracted incorrect	not C&Q	C&Q not against idea	C&Q
#1	27	14	3	10
#2	16	12	0	5
#3	4	2	1	1
#4	3	0	2	1
#5	14	14	0	0
total	64	42	6	17

ances with high precisions and high recalls.

Incorrect utterances extracted by proposed method

Our method, however, extracted some incorrect utterances. We surveyed features of incorrect and extracted utterances. Table 5 shows the features. In this analysis, we decided that correct utterances were defined as utterances which were chosen by more than two participants. The number of features was three. They were all related to utterance styles, criticisms and questions. The first one was that the style was not a criticism or a question. The second one was that the style was a criticism and a question not against idea. The third one was that the style was a criticism and a question. Most of extracted incorrect utterances had styles of the first one: not criticisms or questions. That result indicates that we have set optimistic conditions of judging utterances which styles were criticisms and questions. We have to improve the accuracy of judgments of criticism and question as the future work.

Correct utterances not extracted by proposed method

Our method did not extract some correct utterances. We surveyed features of correct utterances which were not extracted by the proposed method. Table 6 shows the features. In this analysis, we decided that correct utterances were defined as utterances which were chosen by more than two participants. Most of correct utterances not extracted by the proposed method were included in a discussion made by more than three attendees. Our method separate a text into small texts by evaluating utterance exchanges only by two attendees. Our method does not recognize such a discussion as a short text. That result indicates that we need to improve our separating method of text which can detect discussions made by more than three attendees.

6.2 Output by the baseline method and the reason of high accuracy

Example of extracted utterance by the proposed method

Table 7 shows an example of extracted utterance from #2 text by the baseline method. The idea shown in Table 7 is about a system which can move with measuring its present location. The system may be useful for discovering mines under

Table 6: Features and numbers of correct utterances not extracted by proposed. C&Q means a criticism and a question.

Text	Correct but not extracted	More than three attendees	No exchanges	Not C&Q	Short utterance
#1	14	12	1	0	1
#2	19	9	3	7	0
#3	3	1	1	1	0
#4	14	7	2	2	3
#5	12	6	1	5	0
total	62	35	8	15	4

Table 7: Example of extracted utterance by baseline. *Italic font utterance* is the extracted utterance by baseline method

Attendee	Utterance
E	I see. Your system can be used for saving living people.
F	Exactly.
E	<i>Your system could work to help children in Sisen (a city in China) earthquake. To achieve the mission, your system needs GPS sensors. It is quite natural.</i>

the ground and arresting strangers barricading themselves inside. For this idea, the baseline method extracted an utterance *the idea can work in a time of disaster like earthquake. An attendee understood the reason of attaching GPS to the system.* The baseline method evaluates utterances including words with high weights of tfidf. The tfidf method evaluates words appearing frequently in a utterance and words not appearing in other utterances. If the length of utterance is long, a utterance has been prepared by an attendee to be uttered. The utterance includes no new information at least for the attendee. The utterance did not have any potential power for rising up the value of idea. Therefore, the baseline method failed to extract correct utterances.

Incorrect utterances extracted by baseline method

The baseline method extracted some incorrect utterances. We surveyed features of incorrect utterances which were extracted. Table 8 shows the features. In this analysis, we decided that correct utterances were defined as utterances which were chosen by more than two participants. Most of incorrect utterances extracted by the baseline method were idea proposals and solutions for criticisms and questions. The baseline method weights each utterance by weights of words evaluated by tfidf method. Therefore, the baseline method extracts utterances including many words preferentially. These utterances, idea proposals and solutions for criticisms and questions contain many words because many words are needed for explanation. Unfortunately such utterances have less potentials to rise up the value of idea. Therefore, the participants did not choose such utterances as correct utterances.

Table 8: Features and numbers of correct utterances not extracted by baseline method.

Text	# of not extracted	Idea proposal	Added explanation	Solution	C&Q	Remark
#1	50	20	6	7	5	12
#2	95	39	4	39	2	11
#3	29	15	2	12	0	0
#4	18	12	1	5	0	0
#5	70	22	0	42	3	3
Total	262	108	13	105	10	26

Table 9: Averaged tfidf for utterances.

Text	Output	Not output	Correct	Incorrect	All
#1	2676.8	463.4	1311.1	677.7	734.9
#2	5638.8	553.3	769	1255.5	1210.2
#3	6426.2	679.2	1595.9	1368.2	1387.2
#4	7684.9	496.4	1097.3	1139.6	1133.4
#5	4903.1	708.4	1436.9	1287.7	1305.5

Correct utterances not extracted by baseline method

The baseline method did not extract some correct utterances. We surveyed features of correct utterances which were not extracted by the baseline method. Table6 shows the features. The features were about averaged tfidf. In this analysis, we decided that correct utterances were defined as utterances which were chosen by more than two participants. There was no significant difference between the averages of correct utterances and the averages of incorrect utterances. The result indicates that tfidf method fails to extract correct utterances.

7 Conclusion

This paper proposes an extraction method of utterances from a discussion text. The method has developed based on a new discussion model in which criticisms and questions give a topic power to rise up the value of topic.

We experimented with the proposed method, and verified that the proposed method could extract utterances which were useful for rising up the value of idea. The precisions and the recalls in extraction by the proposed method were higher than those by the baseline method using word weights of tfidf.

As the future work, we will try to improve our extraction method of criticism and question, and our separating method of discussion text. We will also try to support reflections of discussion by using the proposed method.

References

- [1] R. Barzilay and M. Elhadad, Using Lexical Chains for Text Summarization, Proc. of ACL on Workshop on Intelligent Scalable Text Summarization, 10–17 (1997).
- [2] G. Erkan and D. R. Radev, LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research 22, 457–479 (2004).
- [3] O. Eris, Effective Inquiry for Innovative Engineering Design, Kluwer Academic Publishers (2004).
- [4] R. A. Finke, T. B. Ward, and S. M. Smith, Creative Cognition Theory, Research, and Application, MIT Press (1992).
- [5] H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, X. Zhang, and G. B. Wise, Cross-document summarization by concept classification, Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 65–69 (2002).
- [6] K. Hori, A System for Aiding Creative Concept Formation, IEEE Transactions on Systems, Man, and Cybernetics, 24(6), 882–894 (1994).
- [7] Y. Gong and X. Liu, Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis, Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (2001).
- [8] K. Knight and D. Marcu, Statistics-Based Summarization — Step One: Sentence Compression, Proc. of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, 703–710 (2000).
- [9] N. Kushiro and Y. Ohsawa, A Requirement Acquisition Process as an Evolved Chance Discovery, Chance Discoveries in Real World Decision Making, Studies in Computational Intelligence, 30, 315–328 (2006).
- [10] H. P. Luhn, A Statistical Approach to the Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development, 1(4), 309 – 317 (1957).
- [11] A. MacLean, T. M. Young, V. Bellotti, and T. Moran, Questions, Options, and Criteria: Elements of design space analysis, Journal of Human-Computer Interaction, 6, 201–250 (1991).
- [12] N. Matsumura, Y. Ohsawa and M. Ishizuka, Mining and Characterizing Opinion Leaders from Threaded Online Discussions, Proc. of the 6th International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, 1267–1270 (2002).
- [13] R. Mihalcea and P. Tarau, A Language Independent Algorithm for Single and Multiple Document Summarization, Proc. of the second International Joint Conference on Natural Language Processing (2005).
- [14] T. P. Moran and J. M. Carroll, Design Rationale: Concepts, Techniques, and Use, Lawrence (1996).
- [15] Y. Nishihara and Y. Ohsawa, Communication Analysis focusing Negative Utterances in Combinatorial Thinking Games, The Review of Socionetwork Strategies, 4(2), 31–46 (2010).
- [16] I. Nonaka and H. Takeuchi, The Knowledge-Creating Company, Oxford University Press (1995).
- [17] A. F. Osborn, Applied imagination: principles and procedures of creative problem-solving, Creative Education Foundation (1953).
- [18] S. Paletz and K. Peng, Implicit Theory of Creating across Cultures-Noveltly and Appropriateness in Two Product Domain, Journal of Cross-Cultural Psychology, Vol.39, pp.286 – 302 (2008).
- [19] B. Pang and L. Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, Proc. of the 42nd ACL, 271–278 (2004).
- [20] K. R. Popper, The Logic of Scientific Discovery, Routledge (1959).
- [21] D. R. Radev, H. Jing, M. Sty and D. Tam, Centroid-based summarization of multiple documents, Information Processing & Management, 40(6), 919–938 (2004).
- [22] G. Salton, Developments in Automatic Text Retrieval, Science, 253, 974–979 (1991).

Enhancing the Identification of Anomalous Events in Medicare Consumer Data Through Classifier Combination

B.Sumudu.U. Mendis[‡], D.Wayne Murray, Alison Sutinen, MingJian Tang, Yingsong Hu

Intelligence Services and Target Identification Portfolio

Department of Human Services, Canberra, Australia

[‡] sumudu.mendis@humanservices.gov.au

Abstract

Fraud detection within Medicare consumer data is a very sensitive issue, especially, when the false positive predictions may adversely affect the relationship between Medicare Australia and the consumer. We here propose a classifier combination model to reduce the risk of false positive predictions in Medicare Australia consumer fraud detection. We also demonstrate our inspiration for the employment of two different computational models which are based on n-Markov model and Frequent Pattern Mining for the identification of anomalous events in consumer sequence patterns. At the end, we discuss the successful outcomes of the classifier combination model.

1 Introduction

In many countries, public health systems are consuming an increasingly significant portion of government expenditure as the population ages and the availability and affordability of medical technology increases. Therefore, it is important for government agencies to continue to monitor their public health system to ensure funding is spent appropriately.

Medicare Australia, a master program within the Department of Human Services, administers the Medicare Benefit Scheme (MBS) which covers medical services and the Pharmaceutical Benefit Scheme (PBS) which covers drug prescriptions. According to its 2009-10 annual report [Medicare Australia, 2010], over 308 million MBS services and over 197 million PBS transactions were processed and paid approximately A\$40 billion in benefits in the 2009-10 financial year. A National Compliance program was developed to outline the key tasks the Department of Human Services will undertake to protect the integrity of its administered programs from the occurrence of inappropriate practice and fraud. The extent of such activities is difficult to quantify but, given the large sums of money involved, even a small percentage can represent a substantial loss of public revenue. According to the National Compliance Program 2010-2011 [Medicare Australia, 2011], “more than \$10.29 million in incorrect payments to providers, pharmacists and members of the public” was identified in 2009-2010. Over the last ten or so years, Medicare Australia has developed, either in-house or through

external engagement, several data mining systems [Pearson *et al.*, 2006; Shan *et al.*, 2008, 2009; Ng *et al.*, 2010] for identifying doctors, pharmacies and consumers that engage in fraudulent or inappropriate billing activities.

This paper expands on the existing work on detecting non-compliant consumers within the health data held by Medicare Australia [Ng *et al.*, 2010]. The previous work focused on consumer fraud/non-compliance in prescription shopping, which refers to the process of consumers targeting doctors and pharmacies to obtain prescription drugs, particularly benzodiazepines and narcotics, in excess of medical needs by deception.

The targeting of fraud by anomalous events¹ in Medicare Australia consumer records involves highly subjective knowledge of MBS, PBS and the current market trends. A single classifier may not be able to model such a complex scenario alone and may not be realistic due to the practical limitations. The nature of some machine learning tasks, eg. Natural language processing (NLP), can be very complex and the classifications must be based on highly subjective knowledge within the domain. In such applications the discovery of combination of multiple classifiers often shows better outcomes than the outcomes based on a single classifier [Brill and Wu, 1998; Florian *et al.*, 2003]. Therefore the combination of several classifiers to model consumer fraud detection could benefit vastly by being able to reduce the false positive outcomes.

In this paper, as a method of classification, we still concentrate on identifying non-compliant consumers by extending the previous work completed in this field [Ng *et al.*, 2010]. Therefore, the first classifier used in this paper builds upon a proposed evaluation framework combining the spatial and temporal nature of the interaction between the consumer, prescriber and the pharmacy. Ng *et al.* [2010] proposed a modified Huff model to model the spatial component of the relationship. In analysing the temporal nature of the prescription drug scripts filled by a consumer, Ng *et al.* [2010] suggested an algorithm based on the Markov model. Please refer to Table 1 in Ng *et al.* [2010] for the complete list of the drugs of interest that used in these studies. Flexibility in this system

¹Identification of an anomalous event is a chance discovery. Chance is a new event conceived either as an opportunity or a risk in the future that may crucially impact on the existing system.

is inferred from the ability to modify this list of drugs as new risks of abuse appear or decline.

Agrawal and Srikant laid the mathematical foundation of the frequent pattern mining for sequence analysis [Agrawal and Srikant, 1995]. Later, PrefixSpan [Pei *et al.*, 2001] algorithm for sequential pattern mining further enhanced the performance. However, all these algorithms focused on finding frequent patterns in sequential data. In this paper, as the second classifier, we propose a novel method called Relative Frequent Pattern Outliers² Index (RFPOI) for ranking outlier-ness of a sequence. At the end, using an experiment we compare the results of the combined classifiers and that of the previous single classifier system.

The paper is organised as follows. Section 2 outlines the existing work undertaken in Medicare Australia to address consumer fraud detection. Section 3 describes the techniques used in this paper to expand the original work. Section 4 looks at the methodology to combine the results of this analysis with the previous analysis to provide the results where each consumer is assigned a rating. The conclusion, future work and discussion is addressed in Section 5.

2 Markov Model for Sequences Outliers Detection

The previous study on which this paper builds, focused on Prescription Shopping, which refers to the process of consumers targeting doctors and pharmacies to obtain prescription drugs in excess of medical needs by deception [Ng *et al.*, 2010]. As such they looked at the three main classes of drugs that are susceptible to abuse by prescription shoppers: opioids, benzodiazepines and psychostimulants ([Ng *et al.*, 2010], Table 1). The probability of an item (i.e. drug) appearing in one episode, is Markovian in depending upon what has come before it in the episode. This is combined with what items (drugs) are in the last episode. Apart from the significant social cost associated with Prescription shopping, there is a financial perspective. The drugs of concern within the Prescription shoppers identified classes of drugs, comprise approximately 6.86% of all PBS prescriptions (in number-of-transactions terms) and cost Medicare Australia approximately \$415 million in the 2008-09 financial year [Ng *et al.*, 2010]. The original paper considered not only the spatial relationship between the consumer and the pharmacy at which the script was filled (using the Huff analysis), but also temporal factors included the PBS item sequences, the sequence of the pharmacies visited by the consumer in each episode of care and the time interval between successive pharmacy visits. The Huff model [Huff, 1964], is a simple stochastic gravity model for studying the effects of distance on consumer choices. This model is trying to quantify the attractiveness to a consumer of a particular pharmacy to fill a script, given the location of the doctor who has written the script.

Using a necessarily small dataset, [Ng *et al.*, 2010] correctly identified 8 of the 12 people as prescription shoppers (see Table 1 in Section 4.3). This paper extends on the ini-

²Outliers is a rare event which significantly differ from other events in the population.

tial work, with its stated objective of trying to reduce the relatively high rate of false positives (approximately 27.78%). We will here describe the method used in [Ng *et al.*, 2010] in brief.

2.1 Huff Model

Ng *et al.* [2010] noted that some compliance officers in Medicare Australia have postulated that some prescription shoppers are willing to travel large distances to different service providers to avoid detection, and that a spatial analysis of consumer data would allow us to detect such behaviour. Taking this into consideration, Ng *et al.* [2010] conceived the need of a measure to quantify spatial behaviour of the consumers. Therefore they propose the Huff model [Huff, 1964] to conduct the spatial analysis. However, in some situations the Huff model may not be directly applicable. Examples of this may include discount pharmacies or pharmacies that open after hours, in which cases the relative attractiveness of different pharmacies needs to be considered.

2.2 n Order Markov Model

The major component of the previous work [Ng *et al.*, 2010] is the calculation of probabilities to analyse the consumer behaviour patterns based on the n-order Markovian model. Now, we here briefly recapitulate the basics of the probabilistic approach they followed.

Their idea is to start with a representative set of item sequences taken from the community. From that they construct a probability distribution over items conditioned on what precedes them and use that to measure how normal a given sequence is with respect to what everybody else in the community is doing [Ng *et al.*, 2010].

Let Σ be a finite set of symbols and Σ^* the set of all finite sequences of symbols in Σ . The analysis of a sequence can be expressed in the following format. For a given a sequence $x_{1:t} = x_1x_2 \dots x_t \in \Sigma^*$, how one can approximate the next symbol x_{t+1} can be calculated by Markovian probability:

$$\Pr(x_{t+1}|x_{1:t}) \quad (1)$$

for all possible $x_{1:t}$ and x_{t+1} . Using Eq.(1), the predictability of a sequence $x_{1:n}$ can be quantified with a measure like:

$$M(x_{1:n}) = -\frac{1}{n} \sum_{i=1}^n \log_2 \Pr(x_i|x_{1:i-1}), \quad (2)$$

where they define $x_{i:j}$, $i < j$. Ng *et al.* [2010] assumes that the suspicious event must take highly predictable PBS item sequences and highly unpredictable pharmacy visits sequences in our PBS database. In section 4, we present the results of this model. The original analysis was undertaken using a combination of the open source statistical software R (for the LOF implementation in the *dprep* package) and in-house C++ programming for the modified Huff model and the temporal feature extraction scheme [Ng *et al.*, 2010]. The main algorithms are all ‘embarrassingly parallel’ in nature and are parallelised using the *OpenMP* package in R.

3 Sequential Pattern Mining For Outliers Detection

Agrawal and Srikant laid the mathematical foundation of the frequent pattern mining for sequence analysis [Agrawal and Srikant, 1995; Srikant and Agrawal, 1996]. Later Srikant and Agrawal generalised their framework and also proposed an improved algorithm called Generalised Sequential Patterns (GAP) algorithm [Srikant and Agrawal, 1996]. GAP is based on the popular Apriori algorithm and hence it also suffers from the disadvantages of Apriori algorithm. One major drawback is that candidate generation can demand high computational complexity as database sizes increase, and this is particularly true when used with Medicare Australia's database. FreeSpan algorithm [Pei *et al.*, 2001] first addresses this issue by using pattern growth approach. Later, PrefixSpan [Pei *et al.*, 2001] algorithm further enhanced the performance of the FreeSpan algorithm. However, all these algorithms focus on finding frequent patterns in sequential data. For example, these algorithms can be directly used to find consumer interests in a supermarket database. Our aim is to modify these methods for mining outliers in sequential data. As the main purpose of the classifier combination is to target outliers, we introduce a novel outlier index called Relative Frequent Pattern Outliers Index (RFPOI). We use the RFPO index system as the second classifier in our classifier combination system to produce the final set of outliers.

3.1 Sequential Pattern Mining

In a nutshell, we can find sequential patterns from a given set of sequences, where each sequence contains a list of itemsets and each itemset contains a set of items. For a given user specified threshold, e.g. minimum support, using a sequential pattern mining algorithm one can find all frequencies of the subsequences that are above or equal to the minimum support. We first demonstrate the frequent pattern mining fundamentals in relevance to the PBS data below.

Let $I = \{i_1, \dots, i_l\}$ be the set of all items. In Medicare perspective I represents the set of all items that are of concern for being abused by some consumers. Let Sequence, $S_i = \langle X_{i,1}, \dots, X_{i,n} \rangle$ s.t. $X_{i,j} \subseteq I$, be an episode of care [Ng *et al.*, 2010] for consumer i .

An item can occur only once in an itemset but it can occur multiple times in a sequence. Also, it is intuitive that the same itemset can occur multiple times in a sequence. Let $S_p = \langle X_{p,1}, \dots, X_{p,k} \rangle$ and $S_q = \langle X_{q,1}, \dots, X_{q,l} \rangle$ be 2 sequences. We can write that S_p is a subset of S_q , ie. $S_p \subseteq S_q$, if there exist integers $1 \leq j_1 < \dots < j_k \leq l$ s.t. $X_{p,1} \subseteq X_{q,j_1}, X_{p,2} \subseteq X_{q,j_2}, \dots, X_{p,k} \subseteq X_{q,j_k}$. Let D be the database of m consumers, $D = \{S_1, \dots, S_m\}$. Now according to [Pei *et al.*, 2001], we can write that the support of a sequence S_α in the sequence database D :

$$Sup_D(S_\alpha) = |\{S_i | (S_i \in D) \wedge (S_\alpha \subseteq S_i)\}| \quad (3)$$

Now, for a given minimum support, say min_sup , one can calculate all the sequences with support greater or equal to the min_sup . Those sequences with equal or higher support than min_sup are called sequential patterns. As we discuss earlier the FreeSpan algorithm improves the performances of

the Apriori based GAP algorithm. However, Pei *et al.* [2001] further improves the FreeSpan algorithm by introducing more efficient algorithm called PrefixSpan. PrefixSpan is based on the fundamental concept that items within an itemset can be in any order. Therefore, it reduces the size of the projected databases used in the FreeSpan algorithm by always listing the itemsets alphabetically. For the work outlined in this paper, we used the PrefixSpan algorithm for the calculation of support values for pharmacy item consumption sequences. The algorithms are implemented in C++ language, *Armadillo* linear algebra package [Sanderson, 2010], and we used OpenMP C++ wrapper package to achieve higher performances by making the algorithm run in parallel. In the next section we explain the fundamentals of our method to find anomalies in sequence databases by extending the frequent sequential pattern mining concept.

3.2 Relative Frequent Pattern Outliers Index (RFPOI)

Our work starts with the basic assumption that anomalous or unusual consumer patterns are rare in a normally distributed sample. Thus, if someone applies sequential pattern mining, those rare occurrences must take less support counts for their itemsets. However it is quite interesting and takes some effort to separate anomalous events from noise and low occurrence customer patterns. Having consulted with the domain experts in Medicare Australia, we speculated that anomalous patterns are quite unique compared to all other patterns. Therefore this work received inspiration from the sequence level support calculation in [Srikant and Agrawal, 1996]. We further extend it to two different support calculations called, *positive support* and *negative support* at the itemset level. We, now extend this idea and present it in a formal mathematical form to define an outliers rank for each sequence based on their consumer behaviour.

We first define the *positive support*, this is the support of any itemset X_α relative to the sequence database D .

Definition 1 Let us take that, $I = \{i_1, \dots, i_l\}$ be the set of all items available and $S_i = \langle X_{i,1}, \dots, X_{i,n} \rangle$ be a sequence i s.t. $X_{i,j} \subseteq I, \forall j = 1, \dots, n$. Let D be the sequence database such that $D = \{S_1, \dots, S_m\}$. The *positive support*, of any itemset X_α relative to the sequence database D is defined as:

$$Sup_D^+(X_\alpha) = |\{X_{i,j} | (X_{i,j} \in D) \wedge (X_\alpha \subseteq X_{i,j})\}| \quad (4)$$

Positive support will measure the appearance of the itemset X_α within the whole database D . Therefore, if the itemset X_α is more common in the database it will show a higher support value and thus, it is less likely to be an outlier. However, during the experiments we discovered that some consumer patterns show higher support values but a large portion of the higher support comes from itself. These consumer patterns look anomalous as they build their own consumer pattern. This further suggests to us the shortfalls of the frequent pattern outlier calculation method shown in [He *et al.*, 2005] to target those consumer patterns. Therefore, we introduce a new concept called *negative support* of an itemset within a sequence to target those patterns. Now, we define

the *negative support*, this is the support of any itemset X_α relative to an arbitrary sequence S_i in database D .

Definition 2 Let us take that, $I = \{i_1, \dots, i_l\}$ be the set of all items available and $S_i = \langle X_{i,1}, \dots, X_{i,n} \rangle$ be a sequence i s.t. $X_{i,j} \subseteq I, \forall j = 1, \dots, n$. Let D be the sequence database such that $D = \{S_1, \dots, S_m\}$. The *negative support* of an itemset X_α relative to the sequence S_i can be calculated as follows:

$$Sup_{S_i}^-(X_\alpha) = |\{X_{i,j} | (X_{i,j} \in S_i) \wedge (X_\alpha \subseteq X_{i,j})\}| \quad (5)$$

Unlike the *positive support*, the *negative support* calculates the support of an itemset X_α within a sequence. Now if a particular consumer is practicing something suspiciously, their consumer behaviour will gradually become different from others and in the end it will show us a unique pattern that can be targeted using *negative support*.

Now we can use a combination of both *positive support* and *negative support* to build the final outliers index for analysing the consumer patterns. We mainly define the outliers index as the ratio between *positive support* and *negative support* multiply by the size of the itemset. The multiplication of *negative support* with the size of the itemset helps us to increase the outlieriness of larger itemsets with high *negative support* compared to that of small itemsets.

Definition 3 For a given minimum support, according to equation (3), let X_α be a sequential pattern. The *Relative Frequent Pattern Outliers Index (RFPOI)* of sequence $S_i \in D$ can be calculated as follows:

$$RFPOI_D(S_i) = \sum_{\forall X_\alpha \subseteq X_{i,j}} \frac{Sup_D^+(X_\alpha)}{Sup_{S_i}^-(X_\alpha) \cdot |X_\alpha|} \quad (6)$$

where $X_{i,j} \in S_i$.

The above definition can be easily extended to calculate the RFPO index for a set of sequences $S_i = \{S_{i,1}, \dots, S_{i,t}\}$. The outcome of the equation (6) is not a normalised value. Therefore, we further extend definition 3 to introduce a normalised RFPO index.

Definition 4 Let $\lambda_i \in \mathbb{R}$. The *normalised Relative Frequent Pattern Outliers Index (RFPOI^N)* can be calculated as follows:

$$RFPOI_D^N(S_i) = \frac{1}{1 + e^{-\lambda_i \cdot RFPOI_D(S_i)}} \quad (7)$$

Where $X_{i,j} \in S_i$.

λ_i in equation (7) is a regularisation parameter to control the normalisation function. Now, any sequence S_i is an outlier if the $RFPOI_D^N(S_i)$ is substantially lower compared other sequence in the database. Later during the experiments we identified 0.001 as the threshold for the value $RFPOI_D^N(S_i)$ to target outliers in the PBS database.

4 Classifier Combination and Experiments

In this section, we first discuss the basics of classifier combination method we used. The later part of this section presents and discusses the experiment and results of the classifier combination. As described in section 3, the classifiers are based on n-Markov model and Frequent Pattern Mining.

4.1 Classifier Combination

As we have shown in the introduction, due to the nature of some problems, the classification could be very complex and involves highly subjective knowledge of the domain. The targeting of fraud by anomalous events in Medicare Australia consumer records involve highly subjective knowledge of MBS, PBS and the current market trends. A single classifier may not be able to model such a complex scenario alone. On the other hand, developing such a single classifier may not be realistic due to the practical limitations. The combination of multiple classifiers [Liao *et al.*, 2010] has become useful in these situations as one classifier may be able to overcome the drawbacks of others and vice versa. Therefore, in our approach we use a combination of different classifiers to model the Medicare consumer fraud detection.

There are many ways to combine the results of the classifiers and to select the final outcome. In classifier combination, it is recommended to use an odd number of classifiers but due to the time limitations we used only 2 classifiers in this study. Results suggest that the classifier combination provided better results which inspired us to consider developing more classifiers in the future. The most straight forward method to combine classifiers is to use equal voting system [Florian *et al.*, 2003; Liao *et al.*, 2010]. This allows the classifiers to have equal opportunity of producing the final results, but in the literature [Liao *et al.*, 2010] it has been shown that in some situations weighted voting provides better results compared to equal voting. However due to time limitations, we investigate the result only from equal voting.

4.2 Experiment

In order to allow direct comparisons between the single classifier method used in [Ng *et al.*, 2010] with the classifier combination method used in this paper, we decided to use the same postcode data. This postcode data represents a major capital city of Australia in which we wish to identify possible fraudulent activities. There are around 30,000 people with registered address in the chosen postcode that have transactional records with Medicare Australia in the original period of study. In the original study, a year worth of transactions have been extracted for each consumer, and those with low volume have been removed, which reduced the number of people down to 15,804 [Ng *et al.*, 2010]. Similar to the original work, we stress that the data does not contain identifying information like consumer names and addresses.

4.3 Results

In the previous work, which is outlined in section 2, the log probabilities of pharmacy item consumption and pharmacy visits have been considered as the two major areas to be measured. In addition to those 2 fields, 5 other fields have also been used for the evaluation of the final outcome of the previous work [Ng *et al.*, 2010]. Interested readers can find the details of those additional measures in [Ng *et al.*, 2010]. In table 1, the 2 columns "LPb(PHI)" and "LPb(PHV)" show the log probabilities of the consumer pharmacy items consumption and the pharmacy visits behaviour (as explained in section 2). In Ng *et al.* [2010], $LPb(PHI) \leq 7$ and

$LPb(PHV) \geq 4.5$ have been used as the 2 threshold values for the pharmacy items consumption and the pharmacy visits log probabilities respectively. As we mentioned earlier the previous system used a total number of seven such flags. Based on these threshold values shown in table 1 row 2, each field triggers a flag for the final evaluation.

Cons. ID	Ent (PRC)	Ent (PH)	LPb (PHI)	LPb (PHV)	Huff Dist.	Old Flags ≥ 3	Expert Prediction
Thres.	> 1.5	> 1.5	≤ 7	≥ 4.5	> 6	yes= 1	yes= 1
1	4	3.5	1.5	6.6	6.8	1	1
2	2.93	2.8	3.7	4.7	5.2	1	1
3	2.39	2.9	12.9	4.8	1.3	1	1
4	1.95	2.9	8	4.6	4.4	1	1
5	1.7	1.8	11.4	2.8	1.1	1	0
6	3.36	2.9	2	5.6	-5	1	1
7	1.78	2.2	8.2	3.2	2.3	1	1
8	2.29	1.9	17.6	2.4	1.6	1	0
9	2.83	2.6	8.9	4.9	0.7	1	1
10	1.59	2.2	16.1	2.4	3.7	1	0
11	1.04	2.1	18.2	6.1	0.9	1	0
12	0.31	2	4.3	3.4	0.2	1	0
13	2.55	2.6	7.2	4.2	3.4	0	1
14	1.09	2.2	3.5	3.9	0.4	0	0
15	1.17	1.1	6.6	1.2	1.1	0	1
16	2.41	2.1	14.5	1.8	4	0	1
17	1.68	1.3	13.6	2.7	0.2	0	0
18	1.82	0.7	14.4	1.6	1.3	0	0

Table 1: Results of the Previous System

According to table 1, the previous system picks out 18 consumers out of 15,804 patients and the Medicare Australia database contains around 220,000 records for those patients. These 18 consumers have at least one flag from the previous system and therefore the domain expert evaluated them all and the results are shown in the column called ‘Expert Prediction’ in table 1. Now, based on the domain experts evaluation, the old system gets 55.56% correct evaluations and 27.78% of false positives and 16.67% of false negatives classifications. Next, we ran the new combined classifier system for the same 15,804 records of data we mentioned above and the results are shown in table 2. Interestingly, the second classifier targets only a subset of the 18 consumers identified by the previous system. The first column, in table 2, shows the enumerated consumer IDs. Each ID refers to the same consumer in both the tables 1 and 2. As we used the previous single classifier system as the first classifier in the new system, we do not repeat the same results in the table 2 as those values are exactly the same as in the table 1. However, in the column called “classif. 1” in table 2, we show the final outcome of the classifier 1 as boolean values, i.e. 1 being a target and 0 indicates a non-target. The columns named “ $RFPOI(PHV)$ ” and “ $RFPOI(PHI)$ ” shows the calculated outliers index values from the second classifier (see section 3) for the consumer pharmacy item (drug) consumption and pharmacy visits patterns. The column called “classif. 2” shows the decisions from the second classifier. The results suggested that, as an individual classifier, the second

classifier shows the best results at the moment. The second classifier, got 72.22% correct evaluations, 16.67% false positive classifications, and 11.11% false negatives (for example Cons. ID 6). Further investigation of false negatives shows that they are quantitatively small and need substantial subjective knowledge to identify. At this point, we do not worry about the false negatives as they are less harmful compared to false positives. However, we expect that by adding more classifiers that could be able to capture the subjective knowledge [Manna *et al.*, 2010] into the system will help us to further reduce the error rate.

Cons. ID	$RFPOI$ (PHV)	$RFPOI$ (PHI)	classif. 1	classif. 2	Comb. classif.	Expert classif.
Thres.	< 0.001	< 0.001	yes= 1	yes= 1	yes= 1	yes= 1
1	1.0000	0.0001	1	1	1	1
2	0.0001	0.0000	1	1	1	1
3	0.0156	0.0002	1	1	1	1
4	0.0095	0.0000	1	1	1	1
5	0.0348	0.0013	1	0	0	0
6	1.0000	1.0000	1	0	0	1
7	0.0187	0.0000	1	1	1	1
8	0.0345	0.0040	1	0	0	0
9	0.0056	0.0001	1	1	1	1
10	0.0836	0.0001	1	1	1	0
11	0.1428	0.0046	1	0	0	0
12	0.0040	0.0001	1	1	1	0
13	0.0095	0.0006	0	1	0	1
14	0.0001	0.0001	0	1	0	0
15	0.0443	0.0002	0	1	0	1
16	0.0394	0.0095	0	0	0	1
17	0.0535	0.0092	0	0	0	0
18	0.0916	0.0011	0	0	0	0

Table 2: Results of the Classifier Combination System

The next column in table 2 (“Comb. classif.”) shows the final outcome of the combined classifier system. It shows not only a good improvement over the previous single classifier system, but also a slightly poorer performance compared to classifier 2 (see section 3) as a single classifier. However the correctness of the results are more reliable compared to single classifiers and we expect that results can be improved by adding more classifiers to the system and also employing more advanced aggregation methods [Mendis *et al.*, 2006] to combine the outcomes of the single classifiers.

5 Conclusions

In this paper we have demonstrated that by extending the work of [Ng *et al.*, 2010] by the inclusion of additional classifiers, we have improved the accuracy of the predicted outcome, whilst also reducing the false positive and false negative rates. The reduction of the false positives is particularly high, for reasons outlined in the abstract. In addition, the paper demonstrated the successful application of the novel RFPO index in targeting outliers.

5.1 Future Work

This work is in its infancy, with small data sets being used

for preliminary evaluation. Having produced encouraging results, the authors now intend to apply the same methodology to a larger dataset.

The finalisation of a Consumer Review Assessment System (Consumer-RAS) would fill the missing link between the consumer, the medical practitioner and the pharmacy³. In this context, the practitioner is evaluated twice, once for their providing of MBS services (the provider) and for the PBS drugs prescribing behaviour (prescriber). Each of these individual entities (consumer, provider, prescriber and pharmacy) are evaluated and rated using a Review Assessment System (RAS) which is a combination of data mining techniques in conjunction with advice from subject matter experts. Thus each entity has a relative weighting as to the fraudulent nature or otherwise of their interactions with Medicare Australia.

There are existing legislative barriers in place in Australia that explicitly prohibit Medicare Australia from linking MBS and PBS data [Ng *et al.*, 2010]. The restriction to link the MBS system with the PBS system would also enforce the separation of these obviously linked analysis.

6 Acknowledgements

The authors wish to thank Dr David Jeacocke, Josephine Ferruga and Ken Yan for their helpful clinical insights. We would also like to thank Lannie Pomazak, Rory King, Dr Jo-Anne Benson and Sheila Bird for providing management support for this paper. We also wish to thank Drs Kee Siong Ng and Yin Shan, both former members of the Medicare Australia's data mining team, for their intellectual leadership and support throughout this process.

References

- R. Agrawal and R. Srikant. Mining sequential patterns. In *icde*, page 3. Published by the IEEE Computer Society, 1995.
- E. Brill and J. Wu. Classifier combination for improved lexical disambiguation. In *Proc. of the 17th int. conf. on comp. linguistics-vol. 1*, pages 191–195. Association for comp. Linguistics, 1998.
- R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. Named entity recognition through classifier combination. In *Proc. of the seventh conf. on Natural language learning at HLT-NAACL 2003-vol. 4*, pages 168–171. Association for comp. Linguistics, 2003.
- Z. He, X. Xu, J.Z. Huang, and S. Deng. Fp-outlier: Frequent pattern based outlier detection. *Computer Science and Information System*, 2(1):103–118, 2005.
- D.L. Huff. Defining and estimating a trading area. *The Journal of Marketing*, 28(3):34–38, 1964.
- ³Future work is underway to apply a modified social network analysis between the different nodes {consumer / (provider/prescriber) / pharmacy}, which can be weighted by the relevant RAS analysis and the edges that link entities can be based upon the number and type of scripts written and filled (to link the consumer with the prescriber and subsequently with the pharmacist) or the number of Medicare services (to link the consumer with the provider).
- J. Liao, B.S.U. Mendis, and S. Manna. Improving hierarchical document signature performance by classifier combination. In *Proc. of the 17th int. conf. on Neural inf. processing: theory and algorithms-Vol. Part I*, pages 695–702. Springer-Verlag, 2010.
- S. Manna, B.S.U. Mendis, and T.D. Gedeon. An enhanced framework of subjective logic for semantic document analysis. *Modeling Decisions for Artificial Intelligence*, pages 231–242, 2010.
- Medicare Australia. Medicare Australia Annual Report 2009-10. In <http://www.medicareaustralia.gov.au/about/governance/reports/09-10/index.jsp>, 2010.
- Medicare Australia. Medicare Australia's National Compliance Program 2010-2011. In <http://www.medicareaustralia.gov.au/provider/business/audits/files/3013-national-compliance-program2010-2011.pdf>, 2011.
- B.S.U. Mendis, TD Gedeon, J. Botzheim, and LT Koczy. Generalised weighted relevance aggregation operators for hierarchical fuzzy signatures. In *Comp. Intell. for Modelling, Control and Automation, and Int. Conf. on Intell. Agents, Web Tech. and Internet Commerce, Int.l Conf. on*, pages 198–198. IEEE, 2006.
- KS Ng, Y. Shan, D.W. Murray, A. Sutinen, B. Schwarz, D. Jeacocke, and J. Farrugia. Detecting Non-compliant Consumers in Spatio-Temporal Health Data: A Case Study from Medicare Australia. In *2010 IEEE Int. Conf. on Data Mining Workshops*, pages 613–622, 2010.
- R. Pearson, D.W. Murray, and T. Mettenmeyer. Finding anomalies in medicare. *Electronic Journal of Health Informatics*, 1(1), 2006.
- J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *icccn*, page 0215. Published by the IEEE Computer Society, 2001.
- C. Sanderson. Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments. In *NICTA Technical Report*, pages 0–15. National ICT Australia, 2010.
- Y. Shan, D. Jeacocke, D.W. Murray, and A. Sutinen. Mining medical specialist billing patterns for health service management. In *Proc. of the conf. in Research and Practice in Inf. Technology*, 2008.
- Y. Shan, D.W. Murray, and A. Sutinen. Discovering inappropriate billings with local density based outlier method. In *Proc. of the conf. in Research and Practice in Inf. Technology*, pages 105–110, 2009.
- R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology—EDBT'96*, pages 1–17, 1996.

Temporal Data Mining in Hospital Information Systems *

Shusaku Tsumoto and Shoji Hirano

Department of Medical Informatics,
Shimane University, School of Medicine
89-1 Enya-cho, Izumo 693-8501 Japan

Abstract

This paper presents application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities are visualized. The results show that the reuse of stored data will give a powerful tool for hospital management and lead to improvement of hospital services.

1 Introduction

Twenty years have passed since clinical data were stored electronically as a hospital information system. Stored data give all the histories of clinical activities in a hospital, including accounting information, laboratory data and electronic patient records. Due to the traceability of all the information, a hospital cannot function without the information system.

However, reuse of the stored data has not yet been discussed in details, except for laboratory data and accounting information to which OLAP methodologies are applied. Data mining approach just started ten years ago [Tsumoto, 2000; 2001].

In this paper, we first propose a scheme for innovation of hospital services based on data mining. Then, based on this scheme, we applied data mining techniques to data extracted from hospital information systems. The results show several interesting results, which suggests that the reuse of stored data will give a powerful tool to improve the quality of hospital services.

The paper is organized as follows. Section 2 proposes a general framework on innovation of hospital services based on data mining. Section 3 briefly explains how hospital information system works, which is a background on this study. Section 3.4 shows the results of visualization of hospital activities by using HIS data. Section 5 provides analysis of clinical process for long-term follow up patients. Section 6 applies trajectories mining technique to temporal analysis the number of orders. Finally, Section 7 concludes this paper.

2 Framework on Data-mining based Hospital Services

Fig 1 shows our goal for hospital services, which consists of the following three layers of hospital management: services for hospital management, services for medical staff and services for patients. Data mining in hospital information system plays a central role in achieving these layers.

The first layer is called services for patients which supports the improvement of healthcare service delivery for patients. This is a fundamental level of healthcare services in which medical staff directly gives medical services to the patients. Patient records and other results of clinical examinations support the quality of this service. The second layer is called services for medical staff which supports decision making of medical practitioner. Patient histories and clinical data are applied to data mining techniques which gives useful patterns for medical practice. Especially, detection of risk of patients, such as drug adverse effects or temporal status of chronic diseases will improve the qualities of medical services. The top layer is called services for hospital management. This level is achieved by capturing global behavior of a hospital: the bridging between microscopic behavior of medical staff and macroscopic behavior of hospital is very important to deploy medical staff in an optimal way for improving performance of the hospital.

3 Background

3.1 Hospital Information System: Cyberspace in Hospital

On the other hand, clinical information have been stored electronically as a hospital information system (HIS). The database stores all the data related with medical actions, including accounting information, laboratory examinations, and patient records described by medical staffs. Incident or accident reports are not exception: they are also stored in HIS as clinical databases. For example, Figure 2 shows the structure of the HIS in Shimane University Hospital. As shown in the figure, all the clinical inputs are shared through the network service in which medical staff can retrieve their information from their terminals [Hanada *et al.*, 2010; Tsumoto and Hirano, 2010]. Since all the clinical data are distributed stored and connected as a large-scale network, HIS can be viewed as a cyberspace in a hospital: all the results

*This research is supported by Grant-in-Aid for Scientific Research (B) 21300052 from Japan Society for the Promotion of Science (JSPS).

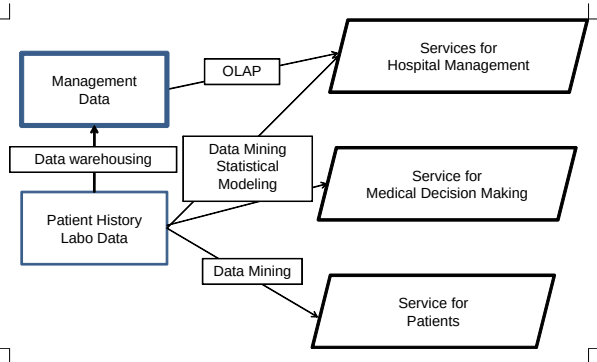


Figure 1: Service-Oriented Hospital Management

of clinical actions are stored as “histories”. It is expected that similar techniques in data mining, web mining or network analysis can be applied to the data. Dealing with cyberspace in a hospital will give a new challenging problem in hospital management in which spatiotemporal data mining, social network analysis and other new data mining methods may play central roles.

3.2 Basic Unit in HIS: Order

The basic unit in HIS is an “order”, which is a kind of document or message which conveys an order from a medical practitioner to others. For example, prescription can be viewed as an order from a doctor to a pharmacist and an prescription order is executed as follows.

1. Outpatient Clinic
2. A prescription given from a doctor to a patient
3. The patient bring it to medical payment department
4. The patient bring it to pharmaceutical department
5. Exceution of order in pharmacist office
6. Delivery of prescribed medication
7. Payment

The second to fourth steps can be viewed as information propagation: thus, if we transmit the prescription through the network, all the departments involved in this order can easily share the ordered information and execute the order immediately. This also means that all the results of the prescription process are stored in HIS.

Figure 3 depicts the workflow of prescription between doctors, pharmacologist, patients and pay desk. For comparison, Figure 4 shows that of injection.

These sharing and storing process, including histories of orders and their results, are automatically collected as a database: HIS can also be viewed as a cyberspace of medical orders.

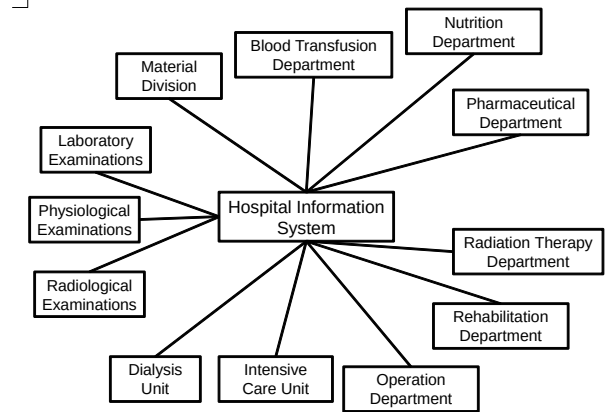


Figure 2: Hospital Information System in Shimane University

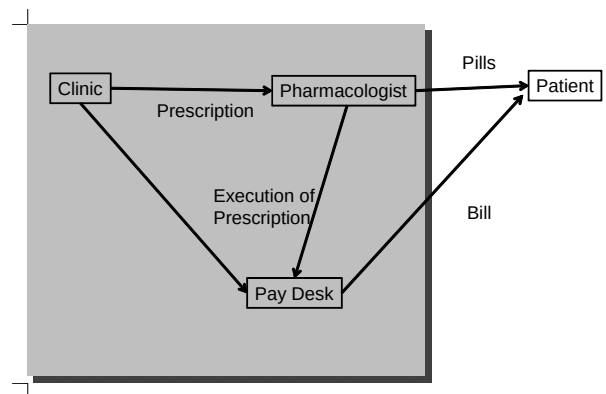


Figure 3: Workflow of Prescription

3.3 Information Reuse Process

Since data in hospital information systems are stored as histories of clinical actions, the raw data should be compiled to those accessible to data mining methods. Although this is usually called “data-warehousing”, medical data-warehousing is different from conventional ones in the following three points. First, since hospital information system consists of distributed and heterogenous data sources. Second, temporal management is important for medical services, so summarization of data should include temporal information. Third, compilation with several levels of granularity is required.

In this paper, we focus on the number of orders to capture temporal global characteristics of clinical activities, whose scheme is given as Fig 5. Here, data-warehousing has two stages: first, we compile the data from heterogenous datasets with a given focus. Then, we count the number of orders with

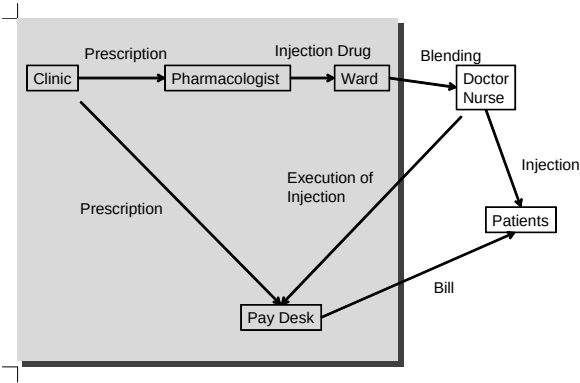


Figure 4: Workflow of Injection

a given temporal section. Data mining process is applied to the secondly generated data sets.

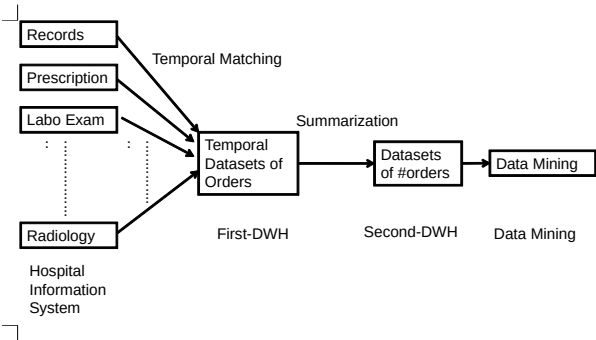


Figure 5: Scheme of Information Reuse

3.4 Visualizing Hospital Actions from Data

Let us show the primitive mining results of HIS. Table 1 shows the averaged number of each order during the same period. Although these values do not remove the effects of holidays, all the characteristics reflect those shown in Figure 2. Patient records and Nursing cares are the major part of orders (39%) and prescription, reservation of clinics, injection are top three orders in the hospital.

Although the above table overview the total behavior of the hospital, we can also check the temporal trend of each order as shown in Figure 6 and 7. The former figure depicts the chronological overview of the number of each order from June 1 to 7, 2008, and the latter shows that of June 2, 2008. Vertical axes denote the averaged number of each order, classified by the type of orders. Horizontal axis give each time zone. The plots show the characteristics of each order. For example, the number of records of doctors has its peak in 11am, which corresponds to the peak of outpatient clinic, whose trend is very similar to reservation of outpatient clinic.

The difference between these two orders is shown in 1pm to 5pm, which corresponds to the activities of wards.

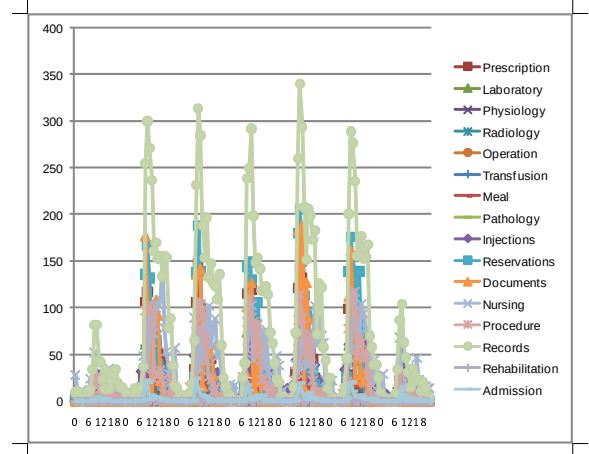


Figure 6: Trends of Number of Orders (June 1 to 6, 2008)

These results show that we can measure and visualize the dynamics of clinical activities in the university hospital by exploratory methods. If we can detect some abnormalities different from the usual behavior in these measurements, this may give some knowledge about risks in the clinical activities. Thus, it is highly expected that data mining methods, especially spatiotemporal data mining techniques play crucial roles in analyzing data in hospital information system and understanding the dynamics of hospital.

4 Clustering of Temporal Trends

Figure 8 shows the grouping of divisions of hospitals by using Ward's method, in which the metrics are calculated from the chronological trends of the number of total orders in outpatient clinic. As shown in this figure, the divisions can be classified into two major categories. Figure 9 shows the results of multidimensional scaling, which gives a two-dimensional complementary view of similarities among divisions.

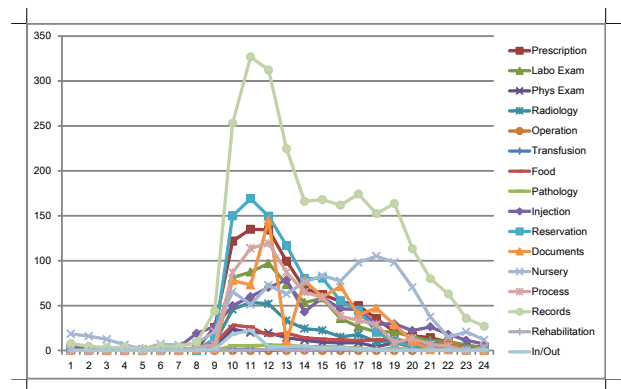


Figure 7: Trends of Number of Orders (June 2, 2008)

Table 1: Averaged Number of Orders per day (Oct, 2006 to Feb, 2011)

	Outpatient			Ward		
	Average	Per Patient	Percentage	Average	Per Patient	Percentage
Prescription	403.0095012	0.45733701	11%	259.8551069	0.626198699	7%
Labo Exam	287.9532858	0.326770695	8%	221.3729216	0.533464349	6%
Phys Exam	78.59540776	0.089190425	2%	45.61995249	0.109934937	1%
Radiology	218.7814727	0.248274208	6%	56.35154394	0.135795921	2%
Operation			0%	12.85273159	0.030972506	0%
Transfusion	1.167854315	0.001325286	0%	11.08709422	0.026717674	0%
Meal			0%	186.2929533	0.44892866	5%
Pathology	21.90894695	0.024862372	1%	16.73713381	0.040333136	0%
Injection	84.41409343	0.095793496	2%	469.3942993	1.131146133	13%
Reservations	663.192399	0.752593743	18%	77.46476643	0.186674553	2%
Documents	454.4845606	0.515751141	13%	156.9659541	0.378256473	5%
Nursery	11.08709422	0.012581685	0%	846.3341251	2.039495526	24%
Process	422.7181314	0.479702453	12%	99.04196358	0.238671271	3%
Records	951.9548694	1.080282704	26%	963.0174188	2.320678865	28%
Rehabilitation	3.047505938	0.003458324	0%	3.064924782	0.007385854	0%
In/Out			0%	55.40300871	0.133510141	2%
Total	3602.315123	4.087923541	100%	3480.855899	8.388164698	100%
#Patients	881.2090261			414.9722882		

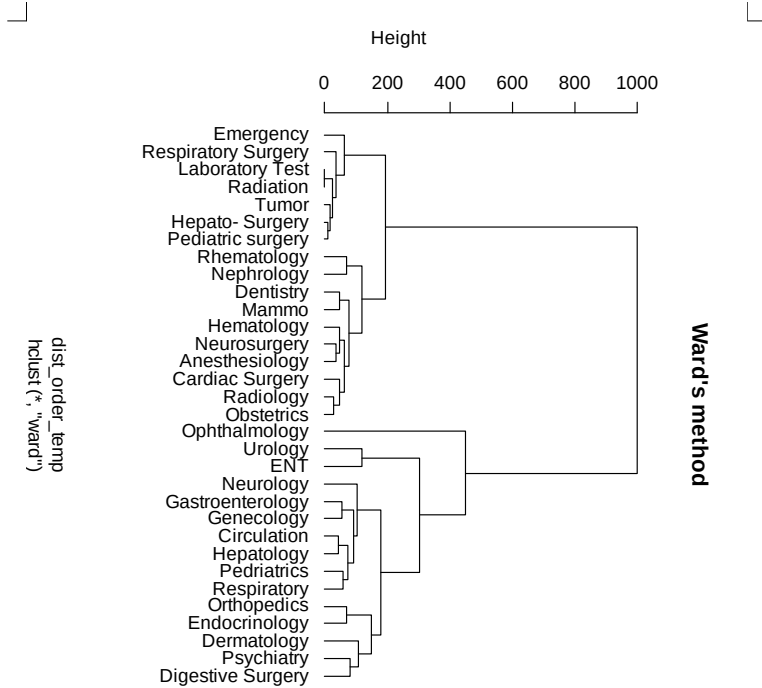


Figure 8: Clustering of Chronological Patterns of # Total Orders

5 Analysis of Clinical Process for Long-term Follow Up Patients

Counting the total number of orders enables us to capture the global nature of clinical process. For example, if a doctor has a patient with more than five years follow up, the total

number of orders for each outpatient clinic gives an overview of its clinical process.

Table 2 shows the statistics of total number of orders for patients who were followed up for more than 5 years. The number of orders per visit is 2.042, and 1.700 of them consists of prescription, laboratory examinations and reservation, all of

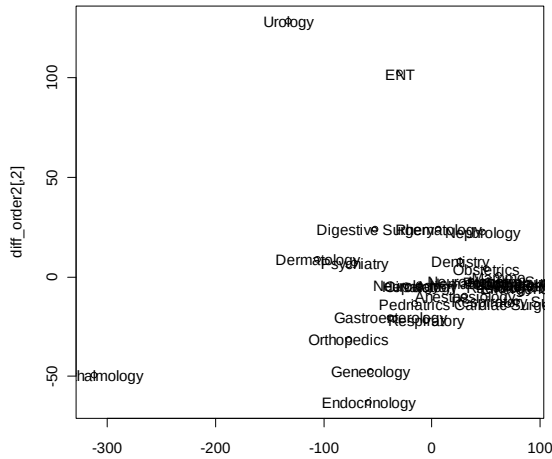


Figure 9: MDS Results of Chronological Patterns of # Total Orders

which are top 3 orders, showing that $1.700/2.042 \sim 83\%$ of clinical orders can be explained by these three orders. Figure 10 shows the decision tree induced for the relations between long-term follow up patients and their visiting clinical departments. This tree also shows that the numbers of three clinical orders characterize those patients. For example, if the number of orders of laboratory examinations is more than 40, 76.2% of the patients are visiting the department of internal medicine.

Table 2: Averaged Number of Orders for Long-Term Follow-up Patients

Orders	Average	#Orders per visit
Total Orders	219.75	2.042
Prescription	90.80	0.845
Labo Exam	51.89	0.483
Reservation	39.81	0.371

6 Analysis of Trajectory of #Orders

If we take two variables of each orders shown in Figure 7, then we can depict the trend of two attributes as a trajectory, as shown in Figures 14 and 15.

Tsumoto and Hirano proposes a clustering method of trajectories, which calculates dissimilarity measures via multi-scale matching and apply clustering methods to trajectories by using the dissimilarities between trajectories [Tsumoto and Hirano, 2009]. By applying this method to the data shown in Subsection 3.4, the dendrogram shown in Figure 11 was obtained. Examples of clusters are shown in Figure 12 and Figure 13. The first one gives a pattern where orders are given both in wards and outpatient clinics. The other one gives a pattern where orders are provided mainly in the wards. A typ-

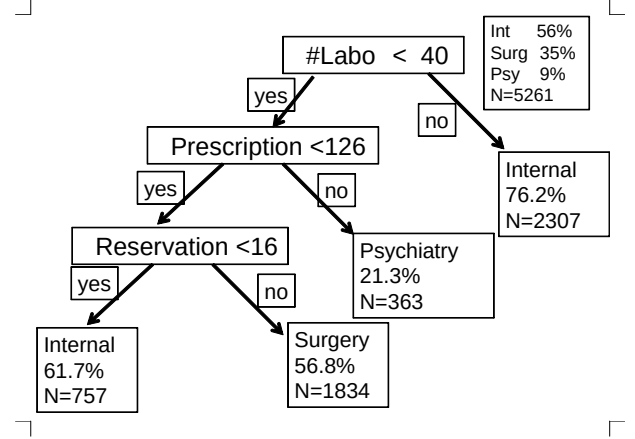


Figure 10: Decision Tree for Long-term Follow-up Patients

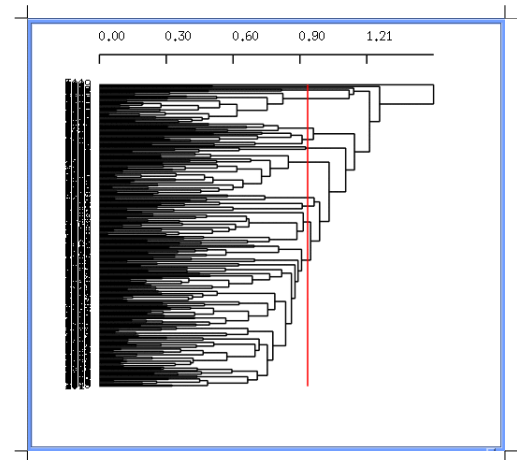


Figure 11: Dendrogram of Trajectories (June 1 to 6, 2008)

ical example in the first cluster is shown in Figure 14, while one in the second cluster is in Figure 15.

7 Conclusions

In this paper, we propose a general framework on innovation of hospital services based on data mining. Then, we applied several data mining techniques to data extracted from HIS in order to capture the characteristics of clinical activities in hospital as follows. First, we shows the chronological overview of hospital activities: periodical behavior of the number of orders can be viewed as a “life-cycle” of a hospital. Secondly, we extracted a pattern of long-term follow up patients with respect to the number of orders by using HIS data. Section 5 provides analysis of clinical process for long-term follow up patients. The number of orders per visit is about 2, and 1.700 of them consists of prescription, laboratory examinations and reservation, all of which are top 3 orders, showing that $1.700/2.042 \sim 83\%$ of clinical orders can be explained by these three orders. Finally, we applied trajectories mining technique to temporal analysis the number of orders. The re-

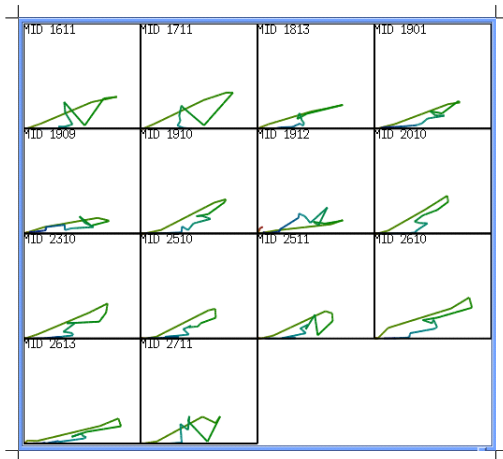


Figure 12: Cluster No.1 (June 1 to 6, 2008)

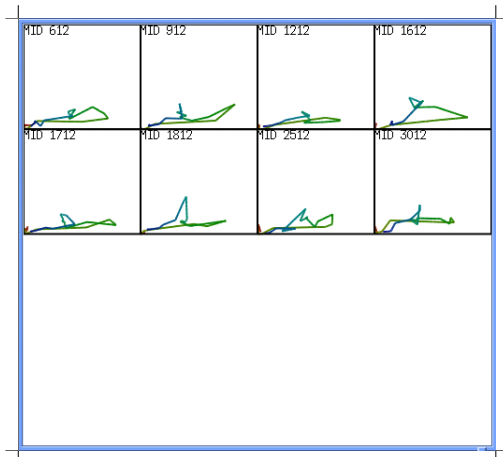


Figure 13: Cluster No.2 (June 1 to 6, 2008)

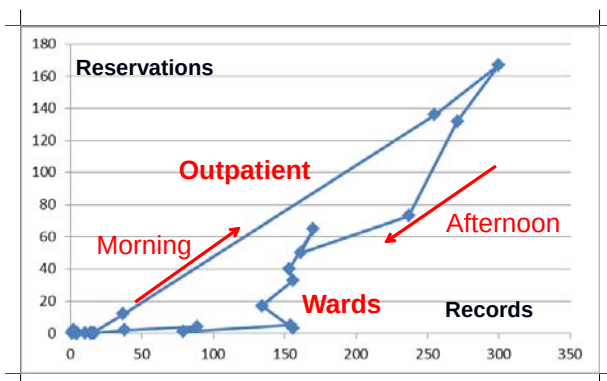


Figure 14: Trajectory between #Reservations and #Records (June 2, 2008)

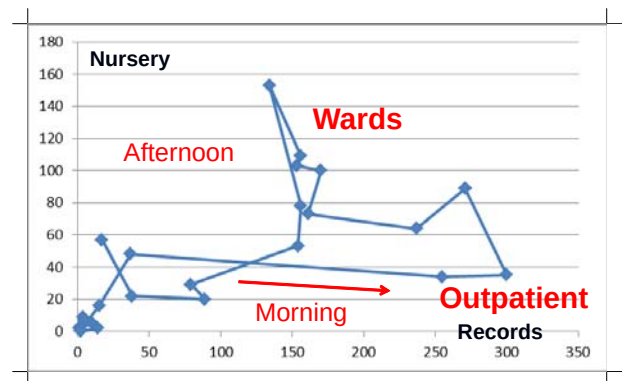


Figure 15: Trajectory between #Nursery Orders and #Records (June 2, 2008)

sults of clustering analysis gave two groups of clinical actions. The one was a pattern where orders are given both in wards and outpatient clinics. The other one was a pattern where orders are provided mainly in the wards.

This paper is a preliminary approach to data-mining hospital management towards a innovative process for hospital services. More detailed analysis will be reported in the near future.

References

- [Hanada *et al.*, 2010] Eisuke Hanada, Shusaku Tsumoto, and Shotai Kobayashi. A ubiquitous environment through wireless voice/data communication and a fully computerized hospital information system in a university hospital. In Hiroshi Takeda, editor, *E-Health*, volume 335 of *IFIP Advances in Information and Communication Technology*, pages 160–168. Springer Boston, 2010.
- [Tsumoto and Hirano, 2009] Shusaku Tsumoto and Shoji Hirano. Detection of risk factors using trajectory mining. *Journal of Intelligent Information Systems*, pages 1–23, 2009. 10.1007/s10844-009-0114-7.
- [Tsumoto and Hirano, 2010] Shusaku Tsumoto and Shoji Hirano. Risk mining in medicine: Application of data mining to medical risk management. *Fundam. Inform.*, 98(1):107–121, 2010.
- [Tsumoto, 2000] S. Tsumoto. Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. *Information Sciences*, (124):125–137, 2000.
- [Tsumoto, 2001] Shusaku Tsumoto. G5: Data mining in medicine. In W. Kloesgen and Jan Zytkow, editors, *Handbook of Data Mining and Knowledge Discovery*, pages 798–807. Oxford University Press, Oxford, 2001.

Discovering Health Problems in the Elderly using Data Mining Approaches

Bogdan Pogorelc, Matjaž Gams

Jožef Stefan Institute & Špica International d.o.o.

Ljubljana, Slovenia

{bogdan.pogorelc, matjaz.gams}@ijs.si

Abstract

Rate of elderly people in the developed countries is larger and larger and not enough younger people exist to take care of them. To provide sustainable elderly care, we propose health care monitoring system, based on data mining approach. The aim of this study is to provide health monitoring system to allow quality and safe living of elderly at homes instead of needing them to go to nursing homes, which are also overcrowded. Moreover, their offspring would not be overwhelmed with care for the elderly. Therefore, the aim of this research is to provide sustainable elderly care. The study proposes a general and specific approach, both achieving classification accuracies over 97%.

1 Introduction

Nowadays, the percentage of elderly people is increasing [Toyne, 2003]. Elderly tend to lead an isolated life away from their offspring; however, they may fear being unable to obtain help if they are injured or ill. During the last decades, this fear has generated research attempts to find assistive technologies for making living of elderly people easier and independent. The aim of this study is to provide ambient assistive living services to allow quality and safe living of elderly at home instead of needing them to go to nursing homes, which are overcrowded. Moreover, their offspring or other relatives would not be overwhelmed with care for the elderly. Therefore, the aim of this research is to provide sustainable elderly care.

We propose two approaches to an intelligent and ubiquitous care system to recognize a few of the most common and important health problems of the elderly, which can be detected by observing and analyzing the characteristics of their movement. In the first approach we use medically defined attributes and support vector machine classification into five health states: healthy, with hemiplegia (usually the result of stroke), with Parkinson's disease, with pain in the leg and with pain in the back [Pogorelc, B. and Gams, M. 2010].

In the second approach we classify into same five health states using more general data mining approach. The movement of the user is captured with the motion capture

system, which consists of the tags attached to the body, whose coordinates are acquired by the sensors situated in the apartment. Output time series of coordinates are modeled with the proposed data mining approaches in order to recognize the specific health problem. In the case that health problem is recognized, the medical center is notified with included explanation of the recognized medical state so that the medical experts can confirm or reject the diagnosis.

2 Related Work

In the related work, motion capturing is usually done with inertial sensors [Strle and Kempe, 2007; Bourke et al, 2006], computer vision and also with specific sensor for measurement of angle of joint deflection [Ribarič and Rozman, 2007] or with electromyography [Trontelj et al, 2008]. For our study, the (infra-red) IR camera system with tags attached to the body [eMotion, 2009] was used.

We do not address the recognition of activities of daily living such as walking, sitting, lying, etc. and detection of falling, which has already been addressed [Confidence, 2009; Luštrek, and Kaluža, 2009], but more challenging recognition of health problems based on motion data.

Using similar motion capture system as in our approach the automatic distinguishing between health problems such as hemiplegia and diplegia is presented [Lakany, 2008]. However, much more common approach to recognition of health problems is capturing of movement which is later examined by medical experts by hand [Ribarič and Rozman, 2007; Craik and Oatis, 1995; Moore et al, 2006]. Such approach has major drawback in comparison to ours, because it needs constant observation from the medical professionals.

The paper [Miskelly, 2001] presented a review of assistive technologies for elderly care. The first technology consists of a set of alarm systems installed at person's homes. A system includes a device in the form of mobile phone, pendant or chainlet that has an alarm button. They are used to alert and communicate with the warden. When the warden is not available, the alert is sent to the control centre. However, such devices are efficient only if the person recognizes an emergency and has the physical and mental capacity to press the alarm button.

The second technology presented in [Miskelly, 2001] is video-monitoring. The audio-video communication is done in real-time over the ordinary telephone line. The video can be viewed on monitor or domestic television. The problems of the presented solution are ethical issues, since the elderly users don't want to be monitored by video [Confidence, 2009]. Moreover, such approach requires constant attention of the emergency center.

The third technology in [Miskelly, 2001] is based on health monitors. The health monitor is worn on the wrist and continuously monitors pulse, skin temperature and movement. At the beginning of the system usage, the pattern for the user is learned. Afterwards, the deviations are detected and alarms are sent to the emergency centre. Such system detects collapses, faints, blackouts etc.

Another presented technology is the group of fall detectors. They measure the accelerations of the person with the tags worn around the waist or the upper chest. If the accelerations exceed a threshold during a time period, an alarm is raised and sent to the community alarm service. Bourke et al. [Bourke et al, 2007] present the acceleration data produced during the activities of daily living and during the person falls. The data was acquired by monitoring young subjects performing simulated falls. In addition, elderly people have performed activities of daily living. By defining the appropriate threshold they can distinguish between the accelerations during the falls and the accelerations produced during normal activities of daily living. Therefore, the accelerometers with the threshold can be used for monitoring elderly people and recognizing falls. However, threshold based algorithms produce mistakes, for instance fast standing up from/sitting down on the chair could result in crossing the threshold which is erroneously recognized as a fall.

In [Rudel, 2008], architecture of a system that enables the control of the users at their homes is described. It consists of three levels. The first level represents the ill persons at their homes equipped with communication and measurement devices. The second level represents information and communication technology that enables the communication with the main server. The last level represents the telemedicine center including duty operator, doctors and technical support; the centre for the implementation of direct assistance at home; and team of experts for implementing telemedicine services. Such system does not provide any automatic detection of an unusual behavior but instead requires constant observation by the medical center.

Williams et al. [2003] have showed that the ability to perform daily activities is decreased for the people that have fallen several times and that the decrease can be detected using accelerometers. They have tested elderly people that have not fallen yet and those that have fallen several times. All of them were asked to perform a predefined scenario including sentence writing, objects picking etc. The accelerations differ significantly between the two groups of people during the test.

The aim of this paper is to realize an automatic classifier able to support autonomous living of elderly by detecting health problems recognizable through the movement. Earlier

works (e.g. [Kaluza et al, 2010]) describe machine learning techniques employed to analyze activities based on the static positions and recognized postures of the users. Although that kind of approaches can leverage a wealth of machine-learning techniques, they fail to keep into account the dynamics of the movement.

3 Materials and Methods

3.1 Targeted Activities and Health Problems for Detection

The research is comparing the specific and the more general approach to recognition of health problems. It classifies walking patterns into five different health states; one healthy and four unhealthy. All the health problems we are recognizing were suggested by the collaborating medical expert on the basis of occurrence in the elderly aged 65+, the medical significance and the feasibility of their recognition from movements.

The following four health problems were chosen as the most appropriate [Craik and Oatis, 1995]:

- Parkinson's disease: a degenerative disease of the brain (central nervous system) that often impairs motor skills, speech, and other functions. The symptoms are frequently tremor, rigidity and postural instability. The rate of the tremor is approximately 4–6 Hz. The tremor is present when the involved part(s), usually the arms or neck, are at rest. It is absent, or diminished with sleep, sedation, and when performing skilled acts.
- Hemiplegia: is the paralysis of the arm, leg and torso on the same side of the body. It is usually the result of a stroke, although diseases affecting the spinal cord and the brain are also capable of producing this state. The paralysis hampers movement, especially walking, and can thus cause falls.
- Pain in the leg: resembles hemiplegia in that the step with one leg is different from the step with the other. In the elderly this usually means pain in the hip or in the knee.
- Pain in the back: There is also a similarity to hemiplegia and pain in the leg in the inequality of steps; however, the inequality is not as pronounced as in walking with pain in the leg.

Classification was done using i) medically defined attributes and SVM classifier and ii) the k-nearest neighbor machine learning algorithm and dynamic time warping for the similarity measure.

3.2 Features for Data Mining

The recordings consisted of the position coordinates for the 12 tags worn on the shoulders, the elbows, the wrists, the hips, the knees and the ankles, sampled with 10 Hz. The tag coordinates were acquired with a Smart IR motion-capture system with a 0.5-mm standard deviation of noise. From the motion capture system we get position of each tag in x-y-z coordinates.

In the first (specific) approach using medically defined attributes 13 attributes were defined with help of an medical expert. These are:

- I. Absolute difference between i) the average distance between the right elbow and the right hip and ii) the average distance between the right wrist and the left hip
- II. Average angle of the right elbow
- III. The quotient between the maximum angle of the left knee and the maximum angle of the right knee
- IV. Difference between the maximum and minimum angle of the right knee
- V. Difference between the maximum and minimum height of the left shoulder
- VI. Difference between the maximum and minimum height of the right shoulder
- VII. Quotient between i) the difference between the maximum and minimum height of the left ankle and ii) the maximum and minimum height of the right ankle
- VIII. Absolute difference between i) the difference between the maximum and minimum speeds (magnitudes of velocity) of the left ankle and ii) the difference between the maximum and minimum speeds of the right ankle
- IX. Absolute difference between i) the average distance between the right shoulder and the right elbow and ii) the average distance between the left shoulder and the right wrist
- X. Average speed (magnitude of velocity) of the right wrist
- XI. Frequency of the angle of the right elbow passing the average angle of the right elbow
- XII. Average angle between i) the vector between the right shoulder and the right hip and ii) the vector between the right shoulder and the right wrist
- XIII. Difference between the average height of the right shoulder and the average height of the left shoulder

We compared the specific approach with the general approach where movements were represented with more general attributes. The advantage of latter approach is that we can add some new health state(s) to be recognized using the same algorithm and attributes.

Considering the abovementioned, in the general approach we designed attributes as the angles between adjacent body parts. The angles between body parts that rotate in more than one direction are expressed with quaternions:

- q'_{SL} and q'_{SR} ... left and right shoulder angles with respect to the upper torso at the time t
- q'_{HL} and q'_{HR} ... left and right hip angles with respect to the lower torso
- q'_T ... the angle between the lower and upper torso
- α'_{EL} , α'_{ER} , α'_{KL} and α'_{KR} ... left and right elbow angles, left and right knee angles.

3.3 Dynamic Time Warping

We will present dynamic time warping (DTW) as a robust technique to measure the “distance” between two time series [Keogh and Ratanamahatana, 2005]. Dynamic Time Warping aligns two time series in the way some distance measure is minimized (usually Euclidean distance is used). Optimal alignment (minimum distance warp path) is obtained by allowing assignment of multiple successive values of one time series to a single value of the other time series and therefore DTW can also be calculated on time series of different lengths. Figure 1 shows examples of two time series and value alignment between them for Euclidean distance (left) and DTW similarity measure (right).

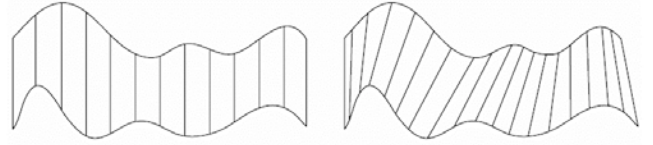


Figure 1: Example of two time series. Lines between time series show value alignment used by Euclidean distance (left) and Dynamic Time Warping similarity measure (right).

Notice that the time series have similar shapes, but are not aligned in time. While Euclidean distance measure does not align time series, DTW does address the problem of time difference. By using DTW, optimal alignment is found among several different warp paths. This can be easily represented if two time series $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_m)$, $a_i, b_j \in \mathbb{R}$ are arranged to form a n -by- m grid. Each grid point corresponds to an alignment between elements $a_i \in A$ and $b_j \in B$. A warp path $W = w_1, w_2, \dots, w_k, \dots, w_K$ is a sequence of grid points, where each w_k corresponds to a point $(i, j)_k$ – warp path W maps elements of sequences A and B .

A warp path is typically subject to several constraints:

- Boundary conditions: $w_1 = (1, 1)$ and $w_K = (n, m)$. This requires the warping path to start in first point of both sequences and end in last point of both sequences.
- Continuity: Let $w_k = (a, b)$ then $w_{k-1} = (a', b')$ where $a - a' \leq 1$ and $b - b' \leq 1$. This restricts the allowable steps in the warping path to adjacent cells.
- Monotonicity: Let $w_k = (a, b)$ then $w_{k-1} = (a', b')$ where $a - a' \geq 0$ and $b - b' \geq 0$. This forces the points in W to be monotonically spaced in time.

From all possible warp paths DTW finds the optimal one:

$$DTW(A, B) = \min_W \left[\sum_{k=1}^K d(w_k) \right]$$

Here $d(w_k)$ is the distance between elements of time series.

Algorithm The goal of DTW is to find minimal distance warp path between two time series. Dynamic programming can be used for this task. Instead of solving the entire problem all at once, solutions to sub problems (sub-series) are found and used to repeatedly find the solution to a slightly larger problem. Let $DTW(A, B)$ be the distance of the optimal warp path between time series $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_m)$ and let $D(i, j) = DTW(A', B')$ be the distance of the optimal warp path between the prefixes of the time series A and B :

$$D(0, 0) = 0$$

$$A' = (a_1, a_2, \dots, a_i), B' = (b_1, b_2, \dots, b_j)$$

$$0 \leq i \leq n, 0 \leq j \leq m$$

Then $DTW(A, B)$ can be calculated using the following recursive equations:

$$D(0, 0) = 0$$

$$D(i, j) = \min(D(i-1, j), D(i, j-1), D(i-1, j-1)) + d(a_i, b_j)$$

Here $d(a_i, b_j)$ is the distance between two values of the two time series (usually Euclidean distance is used). The most common way of calculating $DTW(A, B)$ is to construct a $n*m$ cost matrix M , where each cell corresponds to the distance of the minimal distance warp path between the prefixes of the time series A and B (Figure 2):

$$M(i, j) = D(i, j)$$

$$1 \leq i \leq n$$

$$1 \leq j \leq m$$

We start by calculating all the fields with small indexes and then progressively continue to calculate fields with higher indexes:

```
for i = 1...n
  for j = 1...m
    M(i, j) = min(M(i-1, j), M(i, j-1), M(i-1, j-1)) + dst(ai, bj)
```

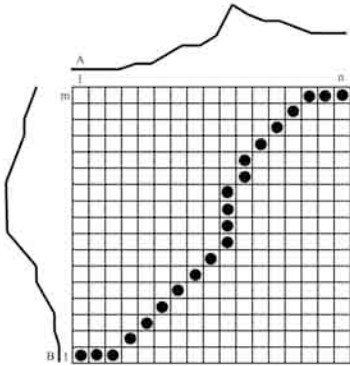


Figure 2: Minimal distance warp path between time series A and B.

The distance corresponding to the minimal distance warp path equals the value in the cell of a matrix M with the highest indexes $M(n, m)$. A minimal distance warp path can be obtained by following cells with the smallest values from $M(n, m)$ to $M(1, 1)$ (in Figure 2 the minimal distance warp path is marked with dots).

Many attempts to speed up the DTWs have been proposed [Salvador and Chan, 2007] which can be categorized as constraints. Constraints limit a minimum distance warp path search space by reducing allowed warp along time axis. Two most commonly used constraints are Sakoe-Chiba Band [Sakoe and Chiba, 1978] which we used and Itakura Parallelogram [Itakura, 1975] which are shown in Figure 3.

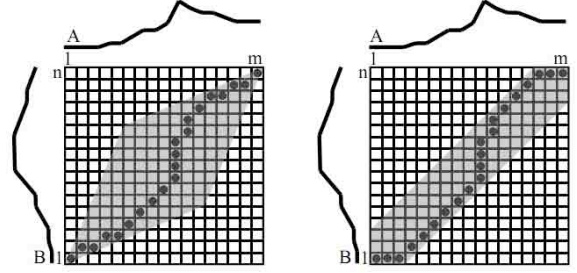


Figure 3: Itakura Parallelogram (left) and Sakoe-Chiba Band (right) constraints. Only shaded cells are used by DTW algorithm.

3.4 Modification of the Algorithm for Multidimensional Classification

The DTW algorithm commonly described in the literature is suitable to align one-dimensional time series. This work employed a modification of the DTW which makes it suitable for multidimensional classification.

First, each time point of the captured time series consisting of the positions of the 12 tags coming out of motion capture system is transformed into angle attribute space, as defined in this paper. The classification will then be performed in the transformed space.

To align an input recording with a template recording (on which the classifier was trained), we first have to compute the matrix of local distances, $d(i, j)$, in which each element (i, j) represents the local distance between the i -th time point of the template and the input at the time j . Let C_{js} be a generic feature vector element relative to a template recording, and Q_{is} be the feature vector element relative to a new input recording to recognize, where $1 \leq s \leq N$ is the considered feature.

For the definition of the local distance the Euclidean distance was used, defined as follows:

$$d_{Euc} = \sum_{s=1}^N (C_{js} - Q_{is})$$

Given the matrix of local distances a matrix of global distances D is built. The value of the minimum global dis-

tance for the complete alignment of DTW procedure, i.e. the final algorithm output, is found in the last column and row, $D(T_r, T_r)$. The optimal alignment can also be efficiently found by back tracing through the matrix: the alignment path starts from $D(T_r, T_r)$, then it proceeds, at each step, by selecting the cell which contains the minimum cumulative distance, between those cells consented by the alignment path constraints, until $D(I, I)$ is reached.

4 Results

In the first (specific) approach, for each recording attributes were calculated and SVM classifier was used to classify them into five health states. Confusion matrix, which represents number of examples of a certain true class (in rows) classified in one of possible five classes (in columns), is shown in Table 1.

In the second (general) approach, the DTW algorithm was used to stretch and compress an input time series in order to minimize a suitably-chosen distance measure from a given template. We used a nearest neighbor classifier based on this distance measure to design the algorithm as a health state classifier.

The classification process is considering one input time series, comparing it with the whole set of templates, computing the minimum global distance for each alignment and assuming that the input recording is in the same class of the template with which the alignment gives the smallest minimum global distance (analogous to instance-based learning). Confusion matrix is shown in Table 2.

The 10-fold cross-validation for 5-nearest neighbor classifier resulted in classification accuracy of 97.9% and 97.6% for the specific and the general approach, respectively. Thus, the performance of both approaches is similar.

For the real world cases, we can use confusion matrices for three purposes:

- False positives (false alarms): How many can be expected using these classifiers. When in real world use the system would report false alarm, e.g., normal walking is classified as some health problem, ambulance could drive to pick up the elderly which would cause unnecessary costs.
- False negatives: How many can be expected using these classifiers. False negatives could mean potentially risky situation for the elderly, as his/her health problem would not be recognized automatically.
- Errors (misclassifications): Between which health states (classes) the errors (misclassifications) occurs. Consequently, we can add additional features to help distinguish between those particular classes. The misclassifications happened very rarely.

		classified as				
		H	L	N	P	B
true class	H	45	0	0	0	0
	L	1	24	0	0	0
	N	0	0	25	0	0
	P	2	0	0	23	0
	B	0	0	0	0	21

Table 1. Confusion matrix for the first (specific) approach, where H=hemiplegia, L=pain in the leg, N=normal (healthy) walking, P=Parkinson's disease and B=Pain in the back. Numbers denote quantity of the classified examples.

		classified as				
		H	L	N	P	B
true class	H	42	2	1	0	0
	L	0	25	0	0	0
	N	1	0	24	0	0
	P	0	0	0	25	0
	B	0	0	0	0	21

Table 2. Confusion matrix for the second (general) approach, where H=hemiplegia, L=pain in the leg, N=normal (healthy) walking, P=Parkinson's disease and B=Pain in the back. Numbers denote quantity of the classified examples.

The results show that in both proposed approaches false positives/negatives are very rare, i.e., they would not cause much unnecessary ambulance costs. Since the method accurately classified most true health problems, it represents high confidence and safety for the potential use in elderly care.

5 Conclusion

This paper proposed elderly health monitoring system providing sustainable elderly care. It presented comparison between specific and general approach to detection of health problems of the elderly. In the specific approach, medically defined attributes and SVM classifier was used. In the general approach, k-nearest neighbor algorithm with multidimensional dynamic time warping was employed. Both approaches classify movement of elderly person into five health states; one healthy and four unhealthy. Even though the first approach is more general and can be used also to classify other types of activities or health problems, it still achieves high classification accuracies, similar to the more specific approach. Since both approaches to health monitoring system are achieving high classification accuracies for recognition of health problems which are also causes for falls, they have advantage over often presented fall detection systems in sense that they recognize health problems in their early stages and also prevent falls.

Acknowledgments

Operation leading to this paper is partially financed by the European Union, European Social Fund. Authors thanks Martin Tomšič, Bojan Nemec and Leon Žlajpah for their help with data acquisition and Anton Gradišek for lending his medical expertise.

References

- [Bourke *et al.*, 2007] Bourke, A. K., O'Brien, J. V., Lyons, G. M. 2007. Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm, *Gait&Posture* 2007; 26:194-99.
- [Bourke *et al.*, 2006] Bourke, A.K. et al. 2006. An optimum accelerometer configuration and simple algorithm for accurately detecting falls. In *Proc. BioMed 2006* (2006), 156–160.
- [Confidence Consortium 2011] Confidence Consortium 2011. Ubiquitous Care System to Support Independent Living. <http://www.confidence-eu.org>.
- [Craik and Oatis, 1995] Craik, R. and Oatis C. 1995. *Gait Analysis: Theory and Application*. Mosby-Year Book (1995).
- [eMotion, 2010] eMotion. 2010. Smart motion capture system. <http://www.emotion3d.com/smart/smart.html>.
- [Itakura, 1975] Itakura, F. 1975. Minimum prediction residual principle applied to speech recognition. *Acoustics, Speech and Signal Processing*, IEEE Transactions on 23(1):67–72.
- [Kaluza *et al.*, 2010] Kaluza, B., Mirchevska, V., Dovgan, E., Lustrek, M., Gams, M., 2010. An Agent-based Approach to Care in Independent Living, *International Joint Conference on Ambient Intelligence (AmI-10)*, Malaga, Spain
- [Keogh and Ratanamahatana, 2005] Keogh, E. and Ratanamahatana, C. A. 2005. "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 358–386, 2005.
- [Lakany, 2008] Lakany, H. 2008. Extracting a diagnostic gait signature. *Patt. recognition* 41(2008), 1627–1637.
- [Luštrek and Kaluža, 2009] Luštrek, M., and Kaluža, B. 2009. Fall detection and activity recognition with machine learning. *Informatica* 33, 2 (2009).
- [Miskelly, 2001] Miskelly, F. G. 2001. Assistive technology in elderly care. *Age and Ageing* 2001; 30:455-58.
- [Moore *et al.*, 2006] Moore, S.T., et al., 2006. Long-term monitoring of gait in Parkinson's disease, *Gait Posture* (2006).
- [Pogorelc and Gams, 2010] Pogorelc, B. and Gams, M. 2010. Identification of Gait Patterns Related to Health Problems of Elderly. *UIC 2010*: 179-191.
- [Ribarič and Rozman, 2007] Ribarič S. and Rozman J. 2007. "Sensors for measurement of tremor type joint movements", *MIDEM* 37(2007)2, pp. 98-104.
- [Rudel, 2008] Rudel, D. 2008. Health at home for elderly. *Infor Med Slov* 2008; 13(2):19-29.
- [Sakoe and Chiba, 1978] Sakoe, H., and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing*, IEEE Transactions on 26(1):43–49.
- [Salvador and Chan, 2007] Salvador, S., and Chan, P. 2007. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* 11(5):561–580.
- [Strle and Kempe, 2007] Strle D., Kempe V., 2007. "MEMS-based inertial systems", *MIDEM* 37(2007)4, pp. 199-209.
- [Toyne, 2003] Toyne, S. 2003. Ageing: Europe's growing problem. BBC News, <http://news.bbc.co.uk/2/hi/business/2248531.stm>.
- [Trontelj *et al.*, 2008] Trontelj, J. et al. 2008. "Safety Margin at mammalian neuromuscular junction – an example of the significance of fine time measurements in neurobiology", *MIDEM* 38(2008)3, 155-160.
- [Williams *et al.*, 2003] Williams, M.E. et al. 2003. A new approach to assessing function in elderly people. *Trans Am Clin Clim Ass* 2003;114:203-16.

Visualizing Systemic Risks and Predicting Systemic Financial Crises using Self-Organizing Maps

Peter Sarlin^a and Tuomas Peltonen^b

^aDepartment of Information Technologies, Åbo Akademi University, Turku, Finland

^bFinancial Stability Surveillance Division, European Central Bank, Frankfurt am Main, Germany

Abstract

The current global financial crisis has demonstrated the importance of understanding sources of domestic and global vulnerabilities that could lead to a systemic event. In this paper, we use the Self-organizing map (SOM) for visualizing systemic risks and predicting systemic events. The SOM model allows disentangling the individual sources of vulnerabilities as well as performs equally well as a logit model in predicting out-of-sample the systemic financial crisis of 2007–09. Further, we introduce the concept of the financial stability cycle and show how the SOM model can be used for monitoring it on a two-dimensional plane.

1 Introduction

The current global financial crisis has demonstrated the importance of understanding sources of domestic and global vulnerabilities that could lead to a systemic event, i.e. an event when the financial instability becomes so widespread that it impairs the functioning of the financial system to the extent that economic growth and welfare suffer materially.¹ The early identification of sources of vulnerabilities are important as they would allow the policymaker to introduce policy actions to decrease or to prevent further building up of vulnerabilities or otherwise enhance the shock absorption capacity of the financial system.

The empirical literature has mostly been based on early warning systems (EWSs) using conventional statistical modeling methods, such as logit and probit models [Berg *et al.*, 2005]. The occurrence of financial crisis is, however, explained by non-linear interactions between non-normally distributed economic and financial variables.² These types of non-linearities derive, for example, from the fact that crises become more likely as the number of fragilities increase. Thus, due to distributional assumptions, conventional statistical modeling techniques may fail in modeling these events. Novel EWSs attempt to model these complex rela-

tionships by applying non-linear techniques. For example, Peltonen [2006] and Fioramanti [2008] show that a neural network outperforms a probit model in predicting currency and debt crises. However, Peltonen [2006] and Berg *et al.* [2005] show that independent of which type of the earlier mentioned models one uses, the results of *a priori* predictions of financial crises are still disappointing. Thus, the utilization of non-linear techniques may have increased *a posteriori* prediction accuracies to a minor extent; however, the interpretability of the monitoring systems has not been addressed appropriately. As a complement to precise numeric predictions, this motivates the development of monitoring systems with clear visual capabilities and intuitive interpretability, contributing instead to tools enabling real human perception.

The dimensionality of the problem complicates interpretability, since a large number of indicators are often required to accurately assess the vulnerability for a financial crisis. Similarly as raw statistical tables, two- and three-dimensional visualizations have limited features for high dimensions. A further challenge is the inclusion of a temporal dimension. Although composite indices of leading indicators, and similarly predicted probabilities as outputs of EWSs, enable comparison across countries and over time, the indices fall short in disentangling the source of vulnerability.³ To this end, a recent work by the IMF staff, called ‘Can you map global financial stability?’ [Dattels *et al.*, 2010], proposes a mapping of six composite indices. However, although emphasizing visual representation of the six indices, the model falls short in disentangling their individual sources. In addition, the authors state that “given the degree of ambiguity and arbitrariness of this exercise the [evaluation] results should be viewed merely illustrative”. Methods for exploratory data analysis (EDA), such as projection and clustering techniques, can contribute to these types of shortcomings. EDA attempts to describe different aspects of the phenomena of interest in an easily understandable form by illustrating the structures in data, but by at the same time preserving information of the original data. The Self-organizing map (SOM) [Kohonen, 2001] is an EDA tool that combines the goals of projection and clustering techniques. It is capable of providing an easily interpret-

¹ Cardarelli *et al.* [2009] show that out of 113 financial stress episodes for 17 main advanced economies, 29 were followed by an economic slowdown and an equal number by recessions.

² Fioramanti [2008], Sarlin and Marghescu [2009] and Lo Duca and Peltonen [2011] show that indicators of debt, currency, and systemic financial crises are non-linearly related.

³ There exist several composite indices for measuring financial tensions, e.g. Illing and Liu [2006] and Cardarelli *et al.* [2009].

able and non-linear description of the multidimensional data distribution without losing the topological relationships of data and sight of individual indicators. The output of the SOM is a mapping of data onto, e.g., a two-dimensional grid, which makes it particularly useful for visualizing large amounts of multidimensional data.

While applications of the SOM concern commonly topics in engineering and medicine, they infrequently focus on financial monitoring. While Arciniegas and Arciniegas Rueda [2009], Resta [2009] and Sarlin and Marghescu [2011] have used the SOM for monitoring indicators of currency crises and Sarlin [2011] has used it for monitoring indicators of debt crises, they solely concentrate on emerging markets. The SOM has not, however, been applied to visualizing and assessing the state of systemic risks and financial stability. The main contribution of this paper is to develop a methodology to visualize the state of financial stability on a two-dimensional plane. The SOM model allows disentangling the individual sources of vulnerabilities, and performs well as an EWS in predicting out-of-sample systemic financial crises.

The paper is structured as follows. Section 2 introduces the SOM, the data and the evaluation framework. In Section 3, we present the training process of the SOM model, evaluate the model and show visual analyses using the model. Section 4 concludes by presenting our key findings.

2 Methodology

There exist two distinguishable categories of EDA methods: projection and clustering methods. The SOM, however, combines the objectives of the EDA groups. The SOM projection differs from other projection techniques, such as multidimensional scaling, by focusing on preserving the neighborhood relations, instead of trying to preserve the distances between data and projects data onto a grid of nodes rather than projecting data into a continuous space. The vector quantization capability of the SOM enables a simultaneous clustering. Further, the two-level clustering of the SOM, i.e., separation of data into nodes and nodes into clusters as proposed in Vesanto and Alhoniemi [2000], enhances the clustering in comparison to other clustering techniques by being more robust non-normal data and efficient and fast and by not needing *a priori* specification of the number of clusters.

2.1 Self-Organizing Maps

The SOM is a two-layered neural network utilizing an unsupervised learning method developed by Kohonen [2001]. It performs a mapping from the input data space $\mathcal{Q}$ onto a k -dimensional array of output nodes. In this paper, we set k equal to 2 for visualization purposes. Vector quantization allows modeling from the continuous space $\mathcal{Q}$, with some probability density function $f(x)$, to a set of nodes. The nodes' location on the map depend on the neighborhood structure of the data in the input space. A second-level clustering can further divide the nodes into clusters. The SOM is only briefly described here, for details see Kohonen [2001].

For its visual representation, the Viscovery SOMine 5.0 package is used in this study. It employs the batch training algorithm, and thus processes data in batches. The most important advantage of the batch algorithm is the reduction of computational cost. The training process starts with an initialization of the reference vectors; they are set as to the two principal components of the input data. PCA initialization not only further reduces computational cost and enables reproducible results, but is also shown to be important for convergence when using the batch SOM [Forte *et al.*, 2002]. Then, the batch training algorithm operates a specified number of iterations according to two steps.

In the first step, each input data vector x is assigned to the most similar reference vector m_c :

$$\|x - m_c\| = \min_i \|x - m_i\|. \quad (1)$$

In the second step, each reference vector m_i (where $i=1,2,\dots,M$) is adjusted using the batch update formula:

$$m_i(t+1) = \sum_{j=1}^N h_{ic(j)}(t) x_j / \sum_{j=1}^N h_{ic(j)}(t). \quad (2)$$

where index j indicates the input data vectors that belong to node c , and N is the number of the data vectors. The neighborhood function $h_{ic(j)} \in [0,1]$ is defined as the following Gaussian function:

$$h_{ic(j)} = \exp\left(-\|r_c - r_i\|^2 / 2\sigma^2(t)\right), \quad (3)$$

where r_c and r_i are two-dimensional coordinates of the reference vectors m_c and m_i , respectively, and the radius of the neighborhood $\sigma(t)$ is a monotonically decreasing function of time t . The radius of the neighborhood begins from being half the diagonal of the grid size ($\sigma = (X^2 + Y^2)^{1/2} / 2$), and goes towards the specified tension value $\sigma(t) \in (0,2]$. The rest of the parameters are the following: map size (the number of nodes), map format (the ratio of X and Y dimensions), and the length of training (training cycles). The second-level clustering is done using an agglomerative hierarchical clustering. The following modified Ward's [1962] criterion is used as a basis for measuring the distance between two candidate clusters:

$$d_{kl} = \begin{cases} \frac{n_k n_l}{n_k + n_l} \cdot \|c_k - c_l\|^2 & \text{if } k \text{ and } l \text{ are adjacent} \\ \infty & \text{otherwise} \end{cases}, \quad (4)$$

where k and l represent clusters, n_k and n_l the number of data points in the clusters and $\|c_k - c_l\|^2$ the squared Euclidean distance between the cluster centers of clusters k and l , and the distance between non-adjacent clusters is infinite.

The output of the SOM algorithm is, for the purpose of this analysis, visualized on a two-dimensional plane. The rationale for using a two-dimensional spectrum, instead of a one-dimensional, is based on the differences within clusters. The multidimensionality of this grid is visualized using layers, namely feature planes. For each individual indicator, a feature plane represents the distribution of its values on the two-dimensional map. In this paper, the feature planes are produced in color, where low to high values are represented by cold to warm scales. The color scales are shown below each corresponding feature plane. Further, the distance be-

tween each node and its corresponding second-level cluster center is indicated with shades; those close to the center take a lighter shade while those further away take a darker shade.

2.2 The Data

The dataset used in this paper is a replica of that in Lo Duca and Peltonen [2011]. It consists of a database of systemic events as well as a set of indicators commonly used in the macroprudential literature to predict financial crises [e.g. Borio and Lowe, 2002; 2004; Alessi and Detken, 2011].

To identify systemic events, we construct a Financial Stress Index (FSI). This approach provides an objective criterion for the definition of the starting date of a systemic event. The larger and broader the shock is (i.e. the more systemic the shock is), the higher the co-movement among variables reflecting tensions. Thus, by aggregating variables measuring stress across market segments, the FSI captures the starting and ending points of a crisis. It is a country-specific composite index, covering the main segments of the domestic financial market and consists of the following five components: (1) the spread of the 3-month interbank rate over the 3-month Government bill rate (Ind_1); (2) negative quarterly equity returns (Ind_2); (3) the realized volatility of the main equity index (as average daily absolute changes over a quarter) (Ind_3); (4) the realized volatility of the nominal effective exchange rate (Ind_4); and (5) the realized volatility of the yield on the 3-month Government bill (Ind_5). Each component j of the index for country i at quarter t is transformed into an integer that ranges from 0 to 3 according to the country-specific quartile of the distribution the observation at quarter t belongs to $q_{j,i,t}$. For example, a value for component j falling into the fourth quartile of the distribution would be transformed into “3”. The FSI is computed for country i at time t as a simple average of the transformed variables as follows:

$$FSI_{i,t} = \sum_{j=1}^5 q_{j,i,t} (Ind_{j,i,t}) / 5 \quad (5)$$

First, the FSI is transformed into a binary variable called “systemic event”. We focus on episodes of extreme financial stress that have often led in the past to negative consequences for the real economy. The variable takes value 1 in the quarter when the FSI moves above the predefined threshold of the 90th percentile of the country distribution. This approach identifies a set of 94 systemic events. Second, we set the class variable C18 to 1 in the 6 quarters preceding the systemic event, and to 0 in all other periods, and define it as the “pre-crisis” period.⁴ The class variable mimics an ideal leading indicator that perfectly signals “systemic events” by “flashing” in the 6 quarters before the event.

We use indicators commonly used in the macroprudential literature to capture the building up of vulnerabilities and

imbalances both in the domestic and global economy. In this regard, we consider equity and property prices and credit developments, valuation levels and proxies for leverage in the economy as key variables. We also include indicators for measuring macroeconomic conditions. To proxy for global macro-financial imbalances and vulnerabilities, we calculate a set of global indicators by averaging the transformed variables for the United States, Euro area, Japan and the United Kingdom. The quarterly dataset consists of 28 countries, of which 10 are advanced and 18 emerging economies, from 1990:1–2010:3. The data sources used are Haver Analytics, Bloomberg and Datastream. As in the literature, several transformations of the indicators, such as annual changes and deviations from moving averages or trends, are constructed to proxy for misalignments and building up of vulnerabilities. The final indicators, and their transformations, are chosen based on their stand-alone performance in predicting systemic events. Statistical properties of the chosen best-performing indicators (Table 1) reveal that the data are significantly skewed and non-mesokurtic, and thus do not exhibit normal distributions.

The analysis is conducted in a real-time fashion to the extent possible. Thus, we take into account publication lags by using lagged variables. For GDP, money and credit related indicators the lag ranges from 1 to 2 quarters depending on the country. We also de-trend variables at time t , using only the information available up to that moment.

2.3 Evaluation Framework

To evaluate the performance of the models in terms of predicting systemic financial crises, we compute a set of accuracy or goodness-of-fit measures. Since we have class information, we use classification performance measures for finding the optimal model instead of the traditional SOM quality measures. We can classify the outcomes into combinations of predicted and actual classes using a contingency matrix. Based on the elements of the matrix, we compute ratios for measuring performance: *recall*, *precision*, *FP* and *FN rates* and *overall accuracy*.⁵ Due to differences in class importance, the above measures are not always appropriate. The costs of giving false signals of crises and tranquil periods might be asymmetric, where the weight depends on the policymaker’s preferences between FNs and FPs.

To calibrate an optimal model and threshold for policy action, we follow the approach first employed in Demirgüç-Kunt and Detragiache [2000] and later enhanced in Bussière and Fratzscher [2008]. The loss function of the policymaker is defined as follows:

$$L(\mu) = \mu(FN/(FN + TP)) + (1 - \mu)(FP/(FP + TN)), \quad (6)$$

where the parameter μ represents the relative preference of the policymaker between FNs and FPs. When $\mu = 0.5$, $\mu > 0.5$ and $\mu < 0.5$, the policymaker is equally, less and more concerned of missing crises and issuing false signals.

⁴ In addition to C18, which is predicted, we set the class variables C24, C18, C12 and C6 to 1 in the 8, 6, 4 and 2 quarters before the systemic event. C0 is the crisis dummy, while P6, P12, P18 and P24 are set to 1 in the 2, 4, 6 and 8 quarters after the systemic event. These are used to assess the financial stability cycle.

⁵ TP = True positives, FP = False positives, TN = True negatives, FN = False negatives, Precision crash = $TP/(TP+FP)$, Recall crash = $TP/(TP+FN)$, Precision tranquil = $TN/(TN+FN)$, Recall tranquil = $TN/(TN+FP)$, Accuracy = $(TP+TN)/(TP+TN+FP+FN)$.

Type	Variable	Abbreviation	Mean	SD	Min.	Max.	Skew.	Kurt.	KSL	AD
Domestic	Inflation ^a	Inflation	0.89	5.17	-10.15	42.53	4.80	26.72	0.29*	263.90*
Domestic	Real GDP ^b	Real GDP growth	3.73	3.76	-17.54	14.13	-0.86	3.16	0.06*	11.34*
Domestic	Real credit to private sector to GDP ^b	Real credit growth	234.07	4724.00	-69.42	101870.34	20.76	429.59	0.51*	Inf*
Domestic	Real equity prices ^b	Real equity growth	5.93	33.01	-84.40	257.04	0.99	4.31	0.05*	7.28*
Domestic	Credit to private sector to GDP ^a	Leverage	3.48	51.64	-62.78	1673.04	22.76	673.35	0.29*	Inf*
Domestic	Stock market capitalisation to GDP ^a	Equity valuation	3.90	28.32	-62.79	201.55	0.77	2.41	0.03*	3.86*
Domestic	Current account deficit to GDP ^c	CA deficit	-0.02	0.07	-0.27	0.10	-0.98	0.73	0.09*	33.12*
Domestic	Government deficit to GDP ^c	Government deficit	0.01	0.05	-0.19	0.22	-1.09	3.46	0.09*	35.90*
Global	Inflation ^a	Global inflation	0.03	0.64	-1.33	2.29	0.71	1.28	0.08*	12.12*
Global	Real GDP ^b	Global real GDP growth	1.84	1.59	-6.34	4.09	-3.02	11.74	0.20*	122.16*
Global	Real credit to private sector to GDP ^b	Global real credit growth	3.87	1.68	-0.23	7.20	-0.21	-0.31	0.07*	8.82*
Global	Real equity prices ^b	Global real equity growth	2.31	19.08	-40.62	37.77	-0.57	-0.68	0.15*	41.90*
Global	Credit to private sector to GDP ^a	Global leverage	1.15	2.79	-2.79	11.21	1.84	3.40	0.22*	105.26*
Global	Stock market capitalisation to GDP ^a	Global equity valuation	0.89	17.41	-40.54	27.46	-0.50	-0.43	0.09*	19.11*

Notes: Transformations: ^a, deviation from trend; ^b, annual change; ^c, level. KSL: Lilliefors' adaption of the Kolmogorov-Smirnov normality test. AD: the standard Anderson-Darling normality test. Significance levels: 1%, *.

Table 1. Statistical properties of the data set

To find out the usefulness of the predictions, we need to subtract the loss for disregarding the information of the model. This is given by $Min(\mu, 1 - \mu)$, i.e., the expected value of a guess with those preferences. Thus, the usefulness of the model is the following:

$$U = Min(\mu, 1 - \mu) - L(\mu). \quad (7)$$

In this paper, we use as a benchmark model a policymaker with $\mu = 0.5$. This could be considered the preferences of a policymaker who does not want to commit any mistakes. Although we do not explicitly assess to what extent policymakers may be more concerned of missing crises than issuing false alarms, we can assume that the occurrence of a crisis is more expensive and should thus weight more.

3 Monitoring the State of Financial Stability

3.1 The SOM Model

Analysis of the financial stability cycle is enabled by introducing class variables for different time periods: pre-crisis (C24, C18, C12, C6), crisis (C0), post-crisis (P6, P12, P18, P24) and tranquil (T0) periods. The pre and post crisis periods range from 24 months before to 24 months after a crisis. In contrast to Sarlin and Marghescu [2011], where the classes are not used in determining the BMU in Eq. 1 (their distribution is only visualized on the map), we let them have a minor impact by giving them a weight of 0.20 when determining the BMUs. Thus, although we give the classes only a minor weight, we employ a semi-supervised SOM by including them in training. This has a cost of lower classification and prediction accuracy, while having a benefit of easier interpretation of the financial stability cycle. To partition the map according to the stages in the financial stability cycle, the nodes of the map are clustered with respect to the class variables using Ward clustering. Further, we obtain the predictive feature of the model by assigning to each data point the C18 value of a node into which the data point is mapped. We evaluate the performance of a model using the usefulness for a policymaker. The performance is computed using static models, i.e. the coefficients or weights are not re-estimated recursively after each point in time. Although static models have the drawback of ignoring the latest available information, they allow better evaluation and comparison

of the models' performance as well as generalization and are necessary for visualizing time series. To test the predictability of the 2008–2009 financial crisis, the train set spans over 1990:4–2005:1, while the test set spans over 2005:2–2009:2. For a parsimonious and objective model, we employ the following training framework.

1. Train and evaluate in terms of in-sample usefulness models for $\sigma \in \{0.0001, 0.3, 0.5, 0.75, 1.0, 1.5, 2.0\}$ and $M \in \{50, 100, 150, 200, 250, 300, 400, 500, 600, 1000\}$.⁶ Set the threshold such that the usefulness is maximized. For each M -value, order the models in a descending order.
2. Find for each M -value the first model with usefulness equal to or better than that of the logit model.
3. Evaluate the interpretability of the models chosen in Step 2. Choose the one that is easiest to interpret.

The chosen model is a map with 137 nodes on an 11x13 grid and trained with a tension of 0.5. Figure 1 presents the two-dimensional SOM grid and feature planes of the 14 indicators as well as of the main class variables. The lines that split the maps into four parts show a clustering of the nodes based on all class variables (Figure 2). In this way, the map is partitioned into four clusters according to the financial stability cycle: a pre-crisis, crisis, post-crisis and tranquil cluster. The upper left cluster represents the pre-crisis cluster (Pre crisis), the lower left represents the crisis cluster (Crisis), the lower right the post-crisis cluster (Post crisis) and the upper right represents the tranquil cluster (Tranquil). The main characteristics of the clusters can be derived from the feature planes in Figures 1 and 2. This enables simultaneous assessment of correlations with all four stages of the financial stability cycle.

3.2 A Comparison with a Logit Model

The logit model estimates are reported in Table 2 and can later be used for predicting out-of-sample data. For both benchmark models ($\mu = 0.5$), the performance is in general similar.

⁶ We keep constant the map format (75:100) and the training length. Kohonen [2001] suggest using a map of oblong shape, while training length is set as an increasing function of map size and decreasing of data points, among other things, by SOMine.

Variable	Estimate	Error	Z	Sig.
Intercept	-6.744	0.612	-11.024	0.000 ***
Inflation	-0.100	0.300	-0.334	0.738
Real GDP growth	0.076	0.334	0.229	0.819
Real credit growth	-0.001	0.001	-0.613	0.540
Real equity growth	1.791	0.382	4.685	0.000 ***
Leverage	0.003	0.001	3.204	0.001 ***
Equity valuation	0.002	0.001	2.689	0.007 ***
CA deficit	1.151	0.308	3.741	0.000 ***
Government deficit	0.076	0.342	0.223	0.823
Global inflation	0.207	0.341	0.608	0.543
Global real GDP growth	1.156	0.419	2.761	0.006 ***
Global real credit growth	0.685	0.381	1.799	0.072 *
Global real equity growth	0.832	0.419	1.985	0.047 **
Global leverage	0.712	0.427	1.668	0.095 *
Global equity valuation	0.959	0.472	2.029	0.042 **

Notes: Significance levels: 1%, ***; 5%, **, 10 %, *.

Table 2. The estimates of the logit model

Model	Data set	μ	Thresh.	Crash periods				Tranquil periods				U	
				TP	FP	TN	FN	Precision	Recall	Precision	Recall		Accuracy
Logit	Train	0.5	0.72	162	190	830	73	0.46	0.69	0.92	0.81	0.79	0.25
SOM	Train	0.5	0.60	190	314	706	45	0.38	0.81	0.94	0.69	0.71	0.25
Logit	Train	0.6	0.54	197	381	639	38	0.34	0.84	0.94	0.63	0.67	0.15
SOM	Train	0.6	0.50	214	419	601	21	0.34	0.91	0.97	0.59	0.65	0.18
Logit	Test	0.5	0.72	77	57	249	93	0.57	0.45	0.73	0.81	0.68	0.13
SOM	Test	0.5	0.60	112	89	217	58	0.56	0.66	0.79	0.71	0.69	0.18
Logit	Test	0.6	0.54	110	109	197	60	0.50	0.65	0.77	0.64	0.64	0.05
SOM	Test	0.6	0.50	134	109	197	36	0.55	0.79	0.85	0.64	0.70	0.13

Notes: TP = True positives, FP = False positives, TN = True negatives, FN = False negatives, Precision crash = TP/(TP+FP), Recall crash = TP/(TP+FN), Precision tranquil = TN/(TN+FN), Recall tranquil = TN/(TN+FP), Accuracy = (TP+TN)/(TP+TN+FP+FN) and usefulness U (formulae 6 and 7).

Table 3. In-sample and out-of-sample performance of the models

On the train set, the SOM model is slightly better than the logit model in terms of usefulness, recall positives and precision negatives, while the logit model outperforms on the other measures. The classification of the models are of opposite nature, as the SOM issues more false alarms (FP rate=31%) than it misses crises (FN rate=19%), while the logit model misses more crises (31%) than it issues false alarms (19%). That explains also the difference in overall accuracy, since the class sizes are unbalanced (80% tranquil periods). The performance of the models on the test set differs, in general, similarly as the performance on the train set. For a policymaker who is less concerned of issuing false

alarms ($\mu = 0.6$), the performance of the models are similar, except for higher usefulness of the SOM model. This shows that the SOM better detects the rare occurrences of crisis.

3.3 Visualizing the State of Financial Stability

In this section, the SOM will be used for two types of visual analyses: for assessing the evolution of a country's macro-financial conditions over time and for benchmarking countries as to their economic conditions on a given point in time

In the middle illustration in Figure 2, the evolution of annual macro-financial conditions are shown for the US and the Euro area from 2002–10. It recognizes for both time series the pre-crisis, crisis and post-crisis stages of the financial stability cycle. Further, the right illustration in Figure 2 shows a mapping of all countries in the third quarter of 2010. This indicates risks in, e.g., Mexico, Turkey, Argentina, Brazil, Taiwan, Malaysia and the Philippines, while most of the advanced countries are either in the post-crisis or tranquil cluster. Singapore, South Africa and India are on the border of the tranquil and pre-crisis clusters which indicates a possible future transition to the pre-crisis cluster.

4 Conclusions

We use the SOM for visualizing and predicting systemic financial crises. The SOM model allows disentangling the individual sources of vulnerabilities and performs equally well as a logit model in predicting the financial crisis of 2007–09. Further, we show how the SOM model can be used for monitoring the financial stability cycle on a two-dimensional plane. Future work should focus on transition probabilities and membership degrees between nodes.

Acknowledgments

The views in this paper are those of the authors and do not necessarily reflect those of the European Central Bank or the Eurosystem.

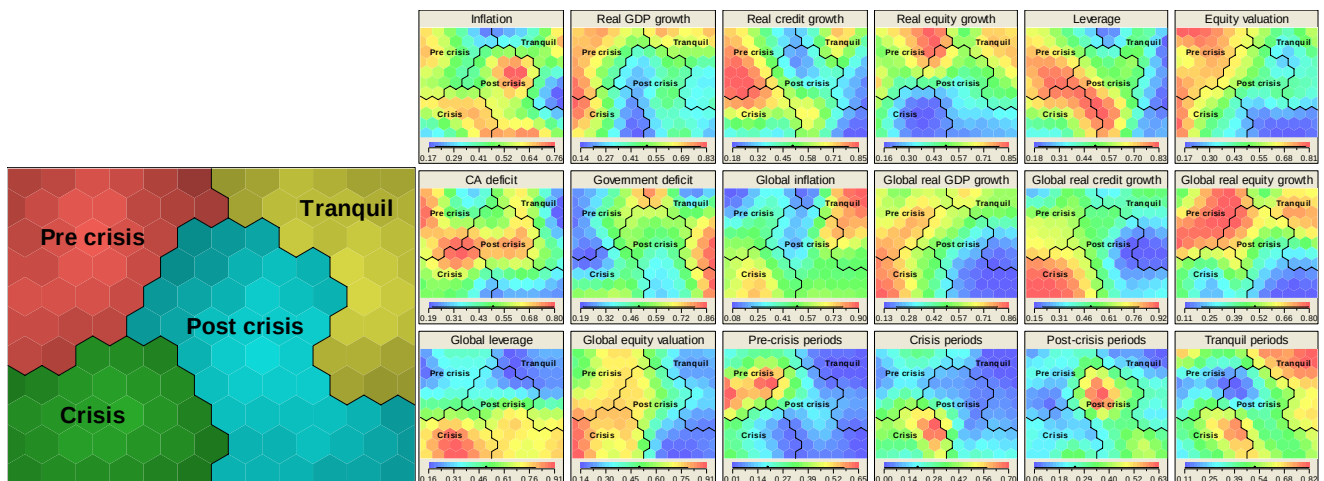


Figure 1. The two-dimensional SOM grid and the feature planes for the 14 indicators and the main class variables

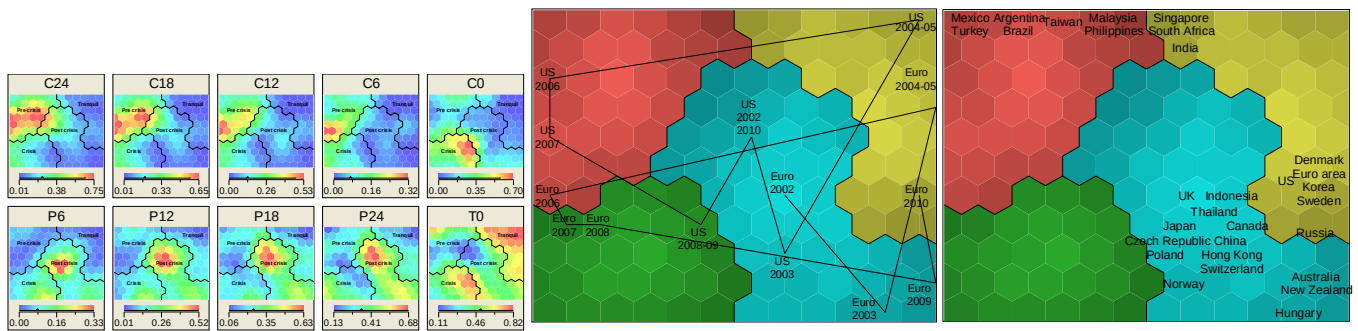


Figure 2. Feature planes of all classes (left), evolution of the US and the Euro area (mid) and a mapping of all countries in 2010Q3 (right).

References

- [Alessi and Detken, 2011] Lucia Alessi and Carsten Detken. Quasi real time early warning indicators for costly asset price boom/bust cycles: a role for global liquidity. *European Journal of Political Economy*, forthcoming.
- [Arciniegas Rueda and Arciniegas, 2009] Ismael Arciniegas Rueda and Fabio Arciniegas. SOM-based data analysis of speculative attacks' real effects. *Intelligent Data Analysis*, 13(2): 261–300, 2009.
- [Berg *et al.*, 2005] Andrew Berg, Eduardo Borensztein and Catherine Pattillo. Assessing early warning systems: How have they worked in practice?. *IMF Staff Papers*, 52: 462–502, 2005.
- [Borio and Lowe, 2002] Claudio Borio and Philip Lowe. Asset Prices, Financial and Monetary Stability: Exploring the Nexus. BIS Working Papers, No. 114, 2002.
- [Borio and Lowe, 2004] Claudio Borio and Philip Lowe. Securing Sustainable Price Stability: Should Credit Come Back from the Wilderness?. BIS Working Papers, No. 157, 2004.
- [Bussiere and Fratzscher, 2008]. Matthieu Bussiere and Marcel Fratzscher. Low probability, high impact: Policy making and extreme events. *Journal of Policy Modeling*, 30: 111–121, 2008.
- [Cardarelli *et al.*, 2009] Roberto Cardarelli, Selim Elekdag and Subir Lall. Financial Stress, Downturns, and Recoveries. IMF Working Paper, WP/09/100, 2009.
- [Dattels *et al.*, 2010] Peter Dattels, Rebecca McCaughrin, Ken Miyajima and Jaume Puig. Can you Map Global Financial Stability?. IMF Working Paper, WP/10/145, 2010.
- [Demirgüç-Kunt and Detragiache, 2000] Asli Demirgüç-Kunt and Enrica Detragiache. Monitoring Banking Sector Fragility: A Multivariate Logit Approach. *World Bank Economic Review*, 14(2): 287–307, 2000.
- [Illing and Liu, 2006] Illing and Liu. Measuring financial stress in a developed country: An application to Canada. *Journal of Financial Stability*, 2: 243–65, 2006.
- [Fioramanti, 2008] Marco Fioramanti. Predicting sovereign debt crises using artificial neural networks: a comparative approach. *Journal of Financial Stability*, 4(2): 149–164, 2008.
- [Forte *et al.*, 2002] Jean-Claude Forte, Patrick Letrémy and Marie Cottrell. Advantages and drawbacks of the Batch Kohonen algorithm. In *Proceedings of the 10th European Symposium on Neural Networks*, pp. 223–230, Bruges, Belgium, April, 2002.
- [Kohonen, 2001] Teuvo Kohonen. *Self-Organizing Maps*, 3rd edition. Springer-Verlag, Berlin, 2001.
- [Lo Duca and Peltonen 2011] Marco Lo Duca and Tuomas Peltonen. Macro-Financial Vulnerabilities and Future Financial Stress – Assessing Systemic Risks and Predicting Systemic Events. ECB Working Paper, No. 1311, 2011.
- [Peltonen, 2006] Tuomas Peltonen. Are emerging market currency crises predictable? A test. ECB Working Paper, No. 571, 2006.
- [Resta, 2009] Marina Resta. Early Warning Systems: an approach via Self Organizing Maps with applications to emergent markets. In *Proceedings of the 18th Italian Workshop on Neural Networks*, pp. 176–184, IOS Press, Amsterdam, Netherlands, 2009.
- [Sarlin, 2011] Peter Sarlin. Sovereign Debt Monitor: A Visual Self-Organizing Maps Approach. In *Proceedings of the IEEE Symposium on Computational Intelligence for Financial Engineering & Economics*. Paris, France, forthcoming, IEEE Press.
- [Sarlin and Marghescu, 2011] Peter Sarlin and Dorina Marghescu. Visual Predictions of Currency Crises using Self-Organizing Maps. *Intelligent Systems in Accounting, Finance and Management*, forthcoming.
- [Ward, 1963] Joe Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58: 236–244, 1963.
- [Vesanto, and Alhoniemi, 2000] Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3): 586–600, 2000.

Map Interface for a Text Data Set by Recursive Clustering

Wataru Sunayama, Shuhei Hamaoka and Kiyoshi Okuda

Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozuka-Higashi, Asa-Minami-Ku, Hiroshima, 731-3194, JAPAN

Abstract

Recently, there are many opportunities to acquire text information as the quantity of electronical information increases. Data classification or clustering methods are widely adapted in order to acquire various information effectively from an enormous active text data set. However, ordinal clustering methods connect texts and many texts are concentrated into a single cluster so that we cannot see various information.

In this study, we propose a recursive clustering method to avoid such bias by integrating a set of texts, included in a cluster, into a single text. An interface that we can comprehend a result of clustering intuitively and can explore information is required to grasp an overview of data and to be led to a new idea. According to the experimental results, the proposed method could construct clusters that are not biased. Test subjects could find information widely by using a map visualizing clustering results.

1 Introduction

Recently, there are many opportunities to acquire text information as the quantity of electronical information increases. Data classification or clustering methods are widely adapted in order to acquire various information effectively from an enormous active text data set. However, ordinal clustering methods connect texts and many texts are concentrated into a single cluster so that we cannot see various information.

Especially, when there are no categories to be classified for an active text data set, clustering methods must focus on its dispersion of clustered data rather than its accuracy. That is, we have a goal to find new ideas by seeing whole data and those meanings, though many ordinal clustering methods considers how classifies data correctly the most.

In this study, we propose a recursive clustering method to avoid such bias by integrating a set of texts, included in a cluster, into a single text. An interface that we can comprehend a result of clustering intuitively and can explore information is required to grasp an overview of data and to be led to a new idea. That is, the following supports are proposed.

1. Improve visibility by visualizing hierarchical data structure constructed by a novel recursive clustering.
2. Display data with a map interface that we can comprehend data intuitively.
3. Display data with a map interface that we can operate to explore data intuitively.

2 Related Works

2.1 Text Classification and Text Clustering

Studies for text classification have been executed since long ago, such as methods with Naive Bayes [1], with EM algorithms [2], with support vector machine [3] and with kernels of feature spaces [4] had been proposed. There are methods to classify texts not only by the original texts but by the labels or tags attached to the texts [5]. Though these methods are available to construct a database automatically, or are used for information retrievals with search engines, more structured clustering results are required to see the whole data and grasp the tendency of a text data set.

2.2 Visualization of Hierarchical Clustering Results

Visualization methods that express hierarchical data structure have been utilized for grasping whole data information and for acquiring information manually. There are methods to visualize a text data set with a hierarchical network, such as a method based on hyperbolic tree [6], data visualization on 3D sphere [7] or on a tree [8], and visualization as a map [9]. Though these methods focused on visualization of classification structure, the clustering methods being employed are major ones like K-means, so those methods do not adopt ideas that changes classification structure. That is, when a system creates clusters for a medium size of a text data set, a set has from 100 to 1000 that we might see all, related to a concrete topic, such text data in a set are not well divided because there are no basic categories to be classified. In such cases, many data are easily categorized in a single cluster and the structure itself become no longer useless.

Therefore, a method that combines a novel clustering method not oriented in its accuracy but in visibility and a map based intuitive interface is proposed in this study.

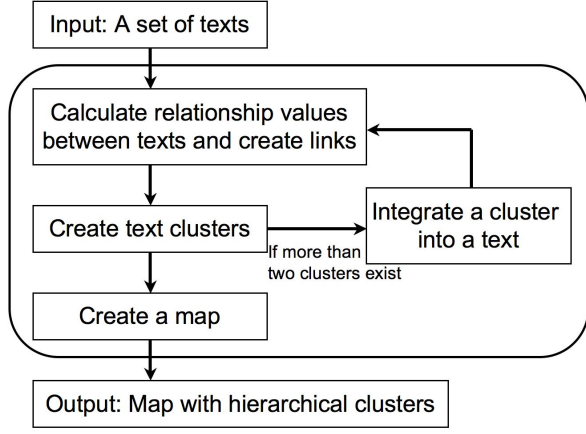


Figure 1: Framework for constructing a map by recursive clustering

3 Map Interface by Recursive Clustering

Fig.1 shows the proposed framework for constructing a map by recursive clustering. The system requires texts as input, the system calculates relationship values between texts at first. Links are created between texts, and clusters are created by the links. If more than two clusters exist, texts in a cluster are integrated into a text and the system creates clusters recursively. Such an iterative process creates a hierarchical clustered structure. Finally, a map is created expressing the constructed hierarchical structure and output.

3.1 Input: A set of texts

Texts including at least one noun can be treated as input. For example, a set of texts obtained as a search result can be an input¹.

3.2 Recursive Clustering

The recursive clustering process consists of three processes, calculation of relationship values between texts for creating links, clustering, and integration of texts in each cluster. Each process is described in the following subsections.

Calculation of Relationship Values and Link Creation

Relationship values between texts are calculated by the cosine similarity given as Eq.(1), where $Num(a)$ denotes the number of words/nouns included in a text a .

$$Relation(a, b) = \frac{Num(a \cup b)}{\sqrt{Num(a) \times Num(b)}} \quad (1)$$

Links are created between texts whose relationship values are more than the threshold K . Though the number of recursive clustering depends of the threshold K , K is set as the maximum value while the number of links do not exceed the half number of texts. For example, when the number of texts

¹The number of texts should be less than 1,000 currently because the algorithm of the process is not well considered. In the future, considering visibility of an output map, 10,000 texts can be available in this environment.

is 100, K is set as the maximum value where the number of links do not exceed 50.

Clustering Method

Newman method [11] is employed as a clustering method. The algorithm is as follows.

- 1) Define each text as a cluster.
- 2) Combine two clusters whose difference of evaluation value Q , that is ΔQ , become the maximum.
- 3) Repeat the step 2) until the value Q become the maximum.

Evaluation value Q is defined as Eq.(2), and its differential value ΔQ is expressed as Eq.(3), where e_{lm} denotes the number of links connecting two texts that belong to clusters l and m each.

$$Q = \sum_l (e_{ll} - a_l^2) \quad (2)$$

$$\Delta Q = e_{lm} + e_{ml} - 2a_l a_m = 2(e_{lm} - a_l a_m) \quad (3)$$

$$a_l = \sum_m a_{lm} \quad (4)$$

This clustering method creates clusters so that the number of links connecting texts belong to the same cluster will increase and the number of links connecting texts that belong to different clusters will decrease.

Integration of Texts in a Cluster

This process integrates texts in each cluster into a text. That is, texts in a cluster are all combined into a text by connecting those texts. It is needless to consider the connecting order because only frequency information is used from each text to calculate relationship.

Recursive Clustering Method

In the recursive clustering process, the amount number of clusters gradually decrease as clusters are integrated. Finally, when the all input texts were integrated into a single text, or when the threshold K became 0.01², the recursive process would be terminated.

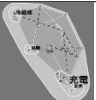
Ordinal many clustering methods tends to create a big cluster that contains many elements especially when a set of texts that are related one another is given. In such cases, the visibility of texts are sacrificed in front of the large mass of data. In the proposed method, even if a large cluster was created once, the texts included in the cluster would be combined into a single text and the weight of the cluster would not become large. On the other hand, this method has a risk to classify texts incorrectly.

3.3 Map Creation

In this section, a drawing method to create a map expressing hierarchical structure is described.

The drawing methods is an extended one of a map animation interface [10]. Primitives in the map are shown in Table 1. The primitives except for “Area” have been already used,

Table 1: Primitives for the proposed map interface

Name	Image	Explanation
Islands		Areas surrounded by the sea, the biggest clusters. An island is a set of areas and texts.
Areas		Areas included in an island, clusters. An area is a set of areas and texts.
Towns		Round nodes. A town is a text.
Roads		Roads are lines connecting towns and areas. The relationship values are more than the threshold.
Rails		Rails are zebra lines connecting towns and areas. The relationship values are more than 0.97.

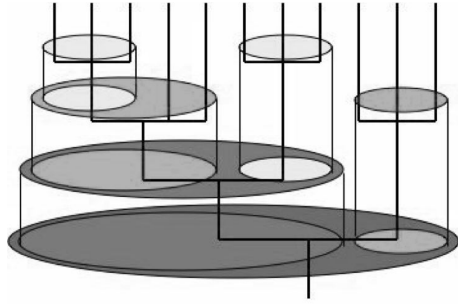


Figure 2: Islands construction by a reversed dendrogram

so the “Area” is newly introduced because “Island” that denotes cluster should be piled up in a hierarchical clustering structure.

Clustering results are expressed as “Islands,” “Rails,” “Roads,” and “Towns” corresponding to clusters, strong links, weak links, and texts respectively when recursive process is not incorporated. Recursive clustering process is expressed as dendrogram, so reversed one like Fig.2 is used for creating a map. That is, each cluster except for the bottom and the second bottom of the Fig.2 called an area. Areas are piled up on islands and islands are on the sea.

The names of islands and areas are extracted from the texts included in those islands or clusters and attached. The names are simply chosen by the document frequency, the number of the documents appearing a specific noun, and the most appeared nouns are extracted as the candidates of labels in the order of frequency. Then, the labels are given to islands at first, and given from the bottom to the top in Fig.2. In this process, if a noun is used as a label once, that noun is not used twice in the direct sequence of the dendrogram, connected as an ancestor and a child.

² K started from 1.00 and decreases by 0.01.

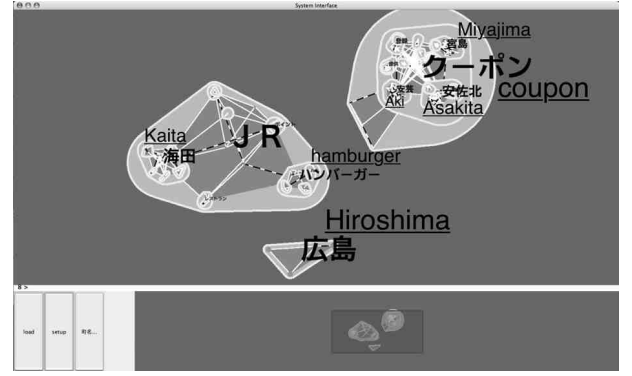


Figure 3: Interface overview

Table 2: Texts used for density evaluation

Themes	Kind of information	Texts
1. Drink	Restaurant information	961
2. Car	Japanese car information	1021
3. World food	Food information	256
4. Game soft	Reviews for games	502
5. Horror	Horrible stories	495
6. Company	Panasonic goods information	626

Fig.3 show an overview of the interface. The background of the map expresses the sea and the map contains three islands that includes piled areas with white lined borders.

3.4 Example Use of Interface

Fig.3 consists of the main map and the sub map that are used for global and local views.

Users basically use the main map using operating commands “Zoom in,” “Zoom out,” “Translation,” and “Selection.” That is, users can use the map like a on-line map service, and “Selection” is realized by double clicks to a town with a mouse to see concrete text information.

For example, suppose a user who wants to have lunch near his university. That user gave texts about shops for lunch as input to the system, and a map is displayed like top left of Fig.4. The user thought that he would like to have lunch near Japanese Railway station, so he zoomed in the island “JR” by using a mouse wheel with translation operation as in the top center of Fig.4. As he saw the “JR” island, he found “Hamburger” area and zoomed in again as in the right of Fig.4. After that, he zoomed in the “Shop” area and double clicked the town “Midorii,” the name of place, and accessed an concrete information like the bottom left of Fig.4.

4 Cluster Density Evaluation

In this chapter, we describe the results of examination whether texts in clusters created by the proposed method are scattered or not.

4.1 Evaluation Method

The examination was executed by comparison between two clustering methods, the proposed method and the minimal

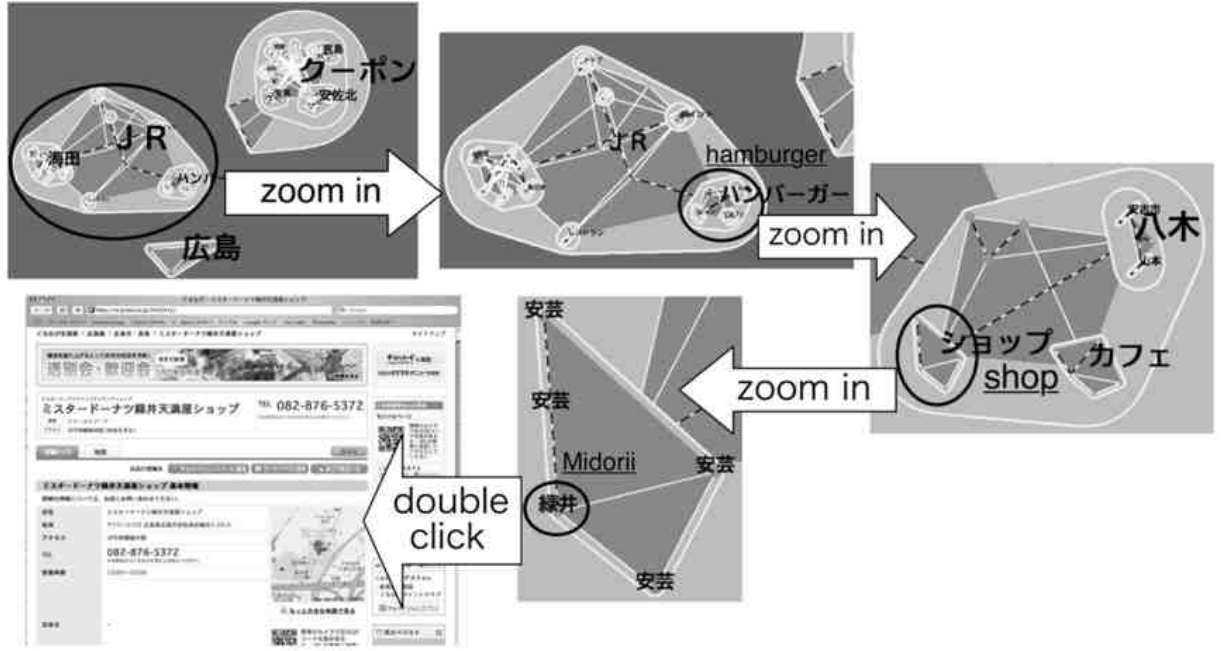


Figure 4: Example operations

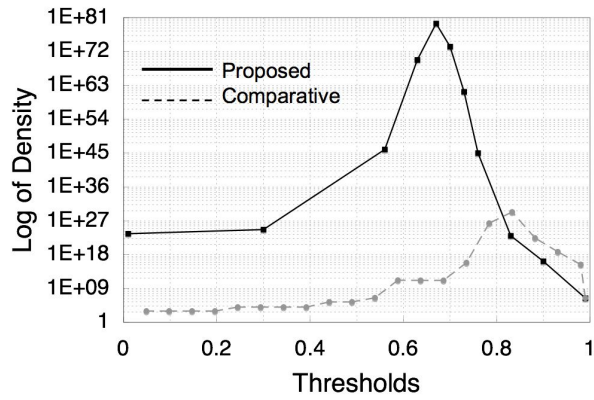


Figure 5: Densities for the theme “Car”

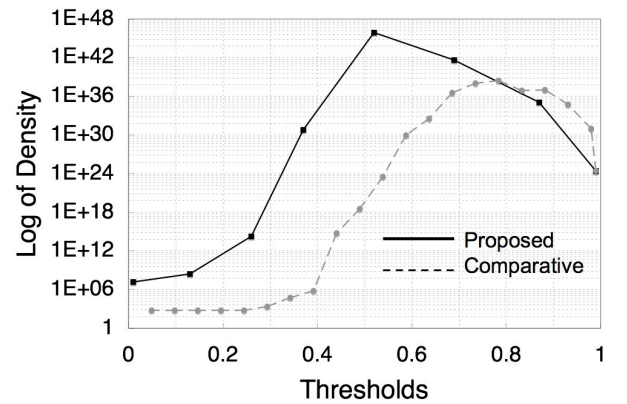


Figure 6: Densities for the theme “Company”

distance method. The minimal distance method combines the most related clusters recursively by treating the relationship value as the most valuable links between two clusters. The data used were shown in Table 2 downloaded from Web sites.

Density value *Density* is defined as Eq.(5) for this examination, where $node(i)$ denotes the number of texts in the cluster i .

$$Density = \prod_i node(i) \quad (5)$$

That is, the density value *Density* become large when the number of data and clusters are the same and the data are scattered equally. In this examination, as the number of data were set as the same, so the method whose value was large could create clusters with no polarization.

4.2 Evaluation Results

Fig.5 and Fig.6 show the results in semi-log graphs. The vertical axis means log of density *Density* and the horizontal axis means the threshold to create links.

According to these figures and results for the other texts, the maximum values in those graphs in the proposed method became very larger than those of the comparative method. Therefore, it is confirmed that the proposed clustering method could disperse data more widely. The values of the left ends were also larger in the proposed method. Since left ends mean the most biggest items, corresponding to islands in the map, so it is expected that we can find various information.

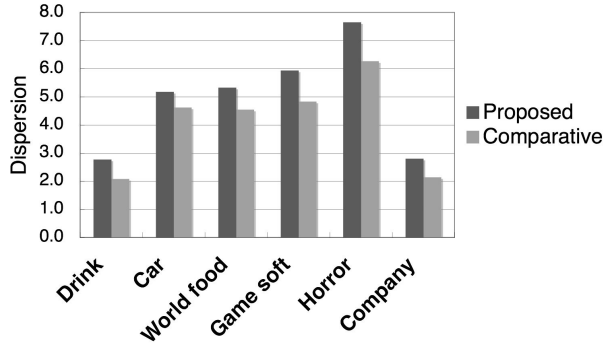


Figure 7: Dispersions for the subjects positively collected

5 Experiments: Information Collection by using Map Interface

In this chapter, the experiments that testify how users can collect information widely with the proposed interface are described.

5.1 Experimental Settings

The texts in Table 2 were used for this experiment. The test subjects were 18 university and graduate school students who majored information sciences. Each test subject collected information according to the given themes such as “You have decided to examine goods whether the company is promising or not. Please pick up goods that express the company well.” in the theme “Company.” That is, themes for information collection were given to the subjects.

The comparative clustering method was the minimal distance method used in the last examination. The map interface was common in two methods. Each subject collected information six times; three times for each. The order of answer and assignment were arranged properly so as not to have biases.

The evaluation should be executed how wide the subjects could collect information. Therefore, *Dispersion* value was defined as the averaged number of meeting points on the dendrogram created by the comparative method. Meeting point is defined as the place where the two texts were combined on the dendrogram. The dispersion value for one meeting point is counted from leaves as zero and increase one by each iterative clustering. That is, the dispersion values become large if collected data were scattered in the dendrogram.

5.2 Experimental Results

Fig.7 shows the dispersion values for the subjects who collected information positively, and Fig 8 shows the dispersion values for the subjects who collected information negatively. “Positive” means the subjects who collected information more than the average in each theme, and “Negative” means the subjects who collected less than the average.

According to the Fig.7, *Dispersions* of the subjects who used the proposed system could more information than those who used the comparative system. This is interpreted that the proposed clustering method scatters texts into many clusters so that the subjects could see various islands or areas.

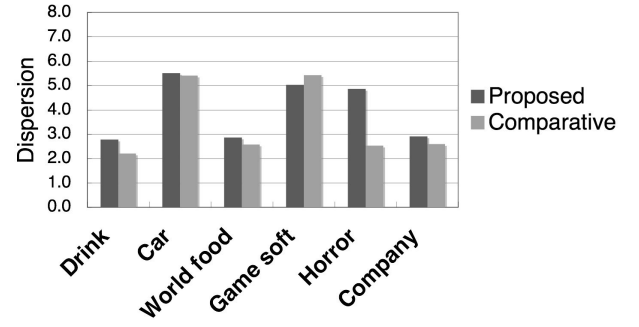


Figure 8: Dispersions for the subjects negatively collected

Table 3: Numbers of collected information

Theme	Proposed	Comparative
Drink	6.0	6.0
Car	6.4	4.2
World food	7.0	5.4
Game	7.0	5.8
Horror	4.3	3.1
Company	5.3	4.9
Average	6.0	4.9

On the contrary, Fig.8, the results for the negative subjects, says that the *Dispersions* were almost the same except for the theme “Horror.” This is because, information was tend to be retrieved from distinctive areas naturally if the quantity of collected information was small. As a result, when a user would like to collect many information, the proposed system is available for making an effective collection.

Table 3 shows the numbers of collected information. The subjects who used the proposed method could collect many information than those with the comparative interface. Therefore, the proposed system could utilized to acquire many and various information.

Table 4 shows concrete information that collected by two subjects who used the proposed system and the comparative one each in the theme “Company.” Both the subjects collected eight information. Though the subject who used the proposed system could collect different types of goods information, two VIERA and three LUMIX are contained in the information collected by the subject who used the comparative system. Therefore, it seems that the comparative method could not divide information widely.

Table 5 shows questionnaire results about visibility and usability. Usability was good in both method because the map interface was intuitive and could be used naturally. However, visibility was not so good in both because the town names and area names were hard to be recognized occasionally. Therefore, the display algorithm should be improved in the future.

6 Conclusion

In this study, a recursive clustering method to improve visibility in order to collect various information is proposed. To

Table 4: Titles of collected informations in the theme “Company”

Proposed method (Dispersion: 2.6)	Comparative method (Dispersion: 1.8)
C3 Liquid crystal television VIERA	Liquid crystal television VIERA
Digital camera DMC-ZX3	X3 Liquid crystal television VIERA
Refrigerator	DMC-TZ LUMIX
Jar pot	DMC-FZ100 LUMIX
Jar Rice cooker	LUMIX Series G, Movie a single lens
Washing machine	Mini compo
Line up for facsimile	Microwaves
Massage chair /	Fitness JOBA
Foot massager	
Buttery	Air purifier

Table 5: Questionnaire results for usability and visibility

	Average (Max : 5 points)	
	Proposed	Comparative
Usability	4.4	4.7
Visibility	3.3	2.7

visualize sophisticated structure is an essential point to grasp the whole data structure and find information to think of new ideas. Clues for new ideas are scattered in enormous information flood, so glittering data for someone are generally hidden. The precious data that are hard to be found do not appear by general visualization methods, so some points such as accuracy that is usually considered as important should be sacrificed in order to find new chances that are not found by anyone.

In future works, the system is expanded to visualize data that should be pointed out by some original measurement scores.

Acknowledgement

A part of this research has been supported by the Kayamori Foundation of Informational Science Advancement.

References

- [1] Andrew McCallum, and Kamal Nigam: A Comparison of Event Models for Naive Bayes Text Classification, *In Proc. of AAAI-98 Workshop on Learning for Text Categorization*, pp.41 – 48, 1998
- [2] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell: Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, Vol.39, pp.103 – 134, 2000.
- [3] Simon Tong, and Daphne Koller: Support Vector Machine Active Learning with Applications to Text Classification, *Journal of Machine Learning Research*, Vol.1, pp.45 – 66, 2001
- [4] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins: Text Classification using String Kernels, *Journal of Machine Learning Research*, Vol.2, pp.419 – 444, 2002

- [5] Grigory Begelman, Philipp Keller, and Frank Smadja: Automated Tag Clustering: Improving search and exploration in the tag space, *In Proc. of Collaborative Web Tagging Workshop at WWW2006*, 2006
- [6] Takumi Kobayashi, Kazuo Misue, Buntarou Shizuki, and Jiro Tanaka: Information Gathering Support Interface by the Overview Presentation of Web Search Results, *In Proc. of Asia-Pacific Symposium on Information Visualisation*, Vol.60, pp.103 – 108, 2006
- [7] T.C.Sprenger, R. Brunella, and M.H.Gross: H-BLOB: A Hierarchical Visual Clustering Method Using Implicit Surfaces, *In Proc. of the 11th IEEE conference on Visualization'00*, 2000
- [8] Mehrnaz Sadat Akhavi, Mohammad Rahmati, and Nerssi Nasiri Amini: 3D Visualization of Hierarchical Clustered Web Search Results, *In Proc. of International Conference on Computer Graphics, Imaging and Visualisation*, pp.441 – 446, 2007
- [9] Markus Muhr, Vedran Sabol, and Michael Granitzer: Scalable Recursive Top-Down Hierarchical Clustering Approach with implicit Model Selection for Textual Data Sets, *IEEE Workshops on Database and Expert Systems Applications*, pp.15 – 19, 2010
- [10] Takuya Nishikido, Wataru Sunayama and Yoko Nishihara: Valuable Change Detection in Keyword Map Animation, *Lecture Notes in Artificial Intelligence 5549*, Springer, pp.233 – 236, 2009
- [11] M.E.Newman: Fast algorithm for detecting community structure in networks, *Phys.Rev. E* 69, 066133, pp. 1 – 5, 2004