2003－ＩＣＳ－133　(9)
2003／9／14

Key Term Extraction and Visualization for Knowledge Chain Discovery Support

－Visual function of TermLinker system－

Konomu DOBASHI[†]　Hiroyuki YAMAUCHI[††]　and　Ryuki TACHIBANA[†††]

[†] Faculty of Modern Chinese Studies, Aichi University　370 Kurozasa, Miyoshi-cho, Nishikamo-gun, Aichi, 470-0296 Japan
[††] Maebashi Institute of Technology　460-1 Kamisadori, Maebashi, Gunma, 371-0816 Japan
[†††] IBM Research, Tokyo Research Laboratory　1623-14 Shimotsuruma, Yamato-shi, Kanagawa, 242-8502 Japan

E-mail:　[†] dobashi@vega.aichi-u.ac.jp

Abstract　This paper presents a method of extracting key terms to visualize the knowledge chain from text corpus. The TermLinker system has developed to extracting cooccurrence relation of key terms from document sentences. The system generates to visualize the knowledge chain using extracted cooccurrence relation of key terms. Proposed method creates partial knowledge hierarchy automatically by concept network. The purpose of the system is to visualize the cooccurrence relationships and help users to understand the relationships of key terms in documents.

Keyword　Text Mining, Key term, Visualization, Knowledge Chain, Discovery Support

キータームの関連性の視覚化による知識連鎖の発見支援

－TermLinkerシステムの可視化機能－

土橋　喜[†]　山内　平行[††]　立花　隆輝[†††]

[†] 愛知大学現代中国学部　〒470-0296 愛知県西加茂郡三好町黒笹 370
[††] 前橋工科大学　〒371-0816 群馬県前橋市上佐鳥町 460-1
[†††] 日本 IBM 東京基礎研究所　〒242-8502 神奈川県大和市下鶴間 1623-14

E-mail:　[†] dobashi@vega.aichi-u.ac.jp

あらまし　本稿では研究者の論文調査における文献収集・内容把握・検索・文章作成などの活動支援を目標に，収集した文献テキストからキータームの共起関係を抽出し，それらから知識連鎖を生成して可視化する方法を提案する．提案する方法はテキストマイニングを応用しており，文章から抽出されたキータームの隣接関係を基に共起関係を生成し，概念ネットワークによって知識連鎖のマップが可視化される．ここで提案する共起関係の生成方法は，概念ネットワークによる知識連鎖のマップの中に，部分的ではあるが階層構造を自動的に生成することができる．システムは知識連鎖の共起関係を可視化することにより，文献に潜在する知識連鎖の発見支援が目的である．

キーワード　キーターム，テキストマイニング，可視化，知識連鎖，発見支援

1. Introduction

Web pages can contain some useful documents, and they are necessary for researcher now. Researchers notice that the documents collected by search engine have relevance to certain themes or keywords.

However, every documents collected by keywords (search words) is not always related to the theme that researchers are looking for. They are just dealing with the same keywords but have nothing with the theme that researchers want. For more effective research, there is the need to tell that the keywords have a relevance to the theme.

Google is one of the search engine to improve this point. It can extract the parts which contain keywords with boldface. Using brief parts, researchers can tell that the keyword is related to the theme. It helps researchers to select at the right documents. To find right documents, researchers need to know about the keywords and other related words, and then check the all of the listed documents. It always happens that the document is different from what researchers expected. Suppose the keywords had thousands of documents and researchers did not know about the keyword much, it would be more difficult to find useful documents for them.

This happens because the conventional search engine cannot figure out the relation between the keywords and other words in the documents and also the importance of the keywords in the document. In addition, the search engines ignore the related words to the keywords in the document, and it decreases the potential to find expected information.

This paper presents a method of extracting key terms to visualize the knowledge chain from HTML documents corpus [9]. The TermLinker system has developed to extracting cooccurrence relation of keywords from document sentences. The system generates to visualize the knowledge chain using extracted cooccurrence relation of keywords [4]. Proposed method creates partial knowledge hierarchy automatically by concept network. The purpose of the system is to visualize the cooccurrence relationships and help users to understand the relationships of key terms in documents [1].

2. Key Terms and Related Words

In billions of digital document, it is important to choose the keywords (search words) carefully to get right information. Users usually need to put other words to decrease the amount of search results. In this point, users need to know about some information related to the keywords to type alternative words and additional words. In this case, it would be more difficult to come up with additional words if users try to find some unfamiliar information. Users need to know some related information, in other words related words, to search about the keywords [8].

Our new TermLinker system of researching the digital documents does not need knowledge especially about keywords (or search words) in advance. It can provide users some other related words, brief overview, and help to make the chapter automatically because of the network of knowledge chain. This concept network is the most vital part of the system we provide. Thus, it is essential to focus on the importance of cooccurrence of keywords and related words to make this concept network[6].

2.1. Extracting key terms

Key terms which are dealt with in the system are recognized as technological terms, compounds, and high frequency words. However, key terms can be treated differently in different language because of characteristic of languages. The language like Japanese needs to be done morphological analysis and divided by space to extract key terms. The language like English is the target of the system. To extract keywords from English documents, it is necessary to normalize each word by stemming and dictionary.

(1) Extracting terms

Technical terms are very important as key terms because they express the characteristic of domain. The system we provide includes dictionary which has 13,000 terms from dictionary of global environment problems.

(2) Extracting compounds

Compounds are important as keywords because they explain important ideas or phenomena. The system extracts compounds which are written more than twice in the documents.

(3) Extracting high frequency words

High frequency words are important as key terms because they express writers` theme. The system extracts words which are written more than three times in the documents.

Stop words and too general words are ignored in extracting keywords. The order of extracting keywords is terms, compounds and then high frequency words. The included dictionary of terms is used to extract low frequency words in the first step.

3. Generating cooccurrence relationships

It is the best way to deal with not words but sentence for generating cooccurrence relationship because it demonstrates a link to the words to other words. Thus, the process of creating cooccurrence relationship is to figure out the relationship of key terms and other words in the focused sentence.

TermLinker system we provide creates the cooccurrence relationship by checking sentence from the beginning and establishing a connection of neighboring keywords. For instance, if the system extracts four key terms like AA, BB, CC, DD, it can create three relationships such as (AA-BB), (BB-CC), (CC-DD). It makes possible to visualize the relationships of key terms.

The system distinguishes AA-BB from BB-AA when it counts the frequency of the keywords. In creating map to visualize the relationship, the most frequent relationship comes first.

4. Generating and visualizing the knowledge chain

An ideal map is a concept network which demonstrates a link to high frequency key terms and clarifies the relationships of each key term and its related terms. In addition, it should be interactive so that it informs right suggestion to users. To create a map like this, the way to link the terms and visualize the related terms is very important.

4.1. Improvement of linking method

New system we developed is different from traditional one in the point of linking the terms. With our traditional system, it was hard to draw the relationships of key terms when the number of key terms is large because the system linked to every key word. New system just links to the neighboring keywords, so it creates a clearer map compared to the traditional system.

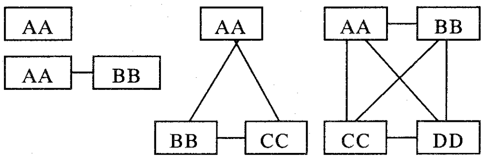

Figure 1 An example of our traditional linking

We defined new cooccurrence rule to connect extracted key terms for mapping the knowledge hierarchy [7]. It's very simple but we think more effective and robust. The process of improving the concept network is just to connect the key terms that TermLilnker system extracts in order. The connections like (AA-BB), (BB-CC) , and (CC-DD) is drew like following figure on the map.

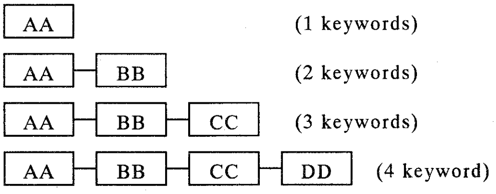

Figure 2 A linking example from a sentence based on new cooccurrence rule

In addition, suppose the system found BB-FF, EE-FF from another sentence, it will draw branched map of which the node is BB (figure 2) .

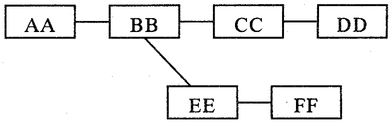

Figure 3 Branched linking

4.2. Visualizing the related key terms

Visualizing the related key terms is the most important part of this system. There are many improved function in the system that we provide. There are two types of visualizing the related key terms in this system [2].

(1) Mapping related key terms users inputted

The first way to display the related key terms is that users type any key terms and system shows the map which is related to the key terms by all of HTML documents database. This is very useful because the system tell the related key terms from database that users have not looked at yet. Users can remind the terms they forgot or did not know because of this map.

(2) Mapping from users selected documents

The second way to display the related key terms is that system creates the map by some documents that user selected. This is useful when user figure out the relationship of key terms from selected documents.

In addition, there are many useful functions to utilize the map. No matter how many key terms user have, user always access to the understandable map because they can modulate the number of relationships which are displayed on the map. The lines of links between key terms display the resource (document) ID and frequency of the relationship of key terms. This link tells users the most important relationships of key terms. If users want to refer original part from documents, users easily access to the sentences from any documents by browse function. It is helpful the review the paper. Highlighting system is also useful. This helps users to understand that the key terms are related to the other terms directly or indirectly. If users click the key terms one time, the most directly related terms are highlighted [3].

5. Experiment

This experiment is for demonstrating the system that we developed. English digital documents from the Internet were appropriate for this experiment. They are published as PDF from Worldwatch Institute. We took 200 documents as sample from following two sites and converted to HTML.

State of the World 2001, pp.298(2001)
Vital Signs 2000, pp.200(2001)
(https://www.worldwatch.org/pubs/).

The collected documents were from 200 words to 17,000 words and their average is 3,258 words (Figure 1). To see the documents by browse, we put HTML tag on top and ending of the documents.

Figure 1 The number of words and cooccurrence relationship（The number of documents：200）

	minimum	max	average
Number of words	218	17,015	3,258
Number of key terms	33	1,183	275
Number of cooccurrence relationship	30	3,200	569

Figure 2 shows the sum of number of words, key terms, and cooccurrence relationship on the map. There is no double counting keywords and cooccurrence relationship.

The first sample we present (Figure 4) is the fundamental cooccurrence map for one document. The title of document is "Paper Recycling Remains Strong". The detail data of map is following.

Figure 2 The number of words, key terms and cooccurrence relationships on the map

	Number of words	Number of keywords	Number of cooccurrence relationships
Figure 4	1,274	304	223

The map become more complicated because there are a lot of terms of which frequency is just one time on the map. This is because researchers rarely state the same expression in the documents. To decrease the number of key terms by frequency is necessary for simple understandable map.

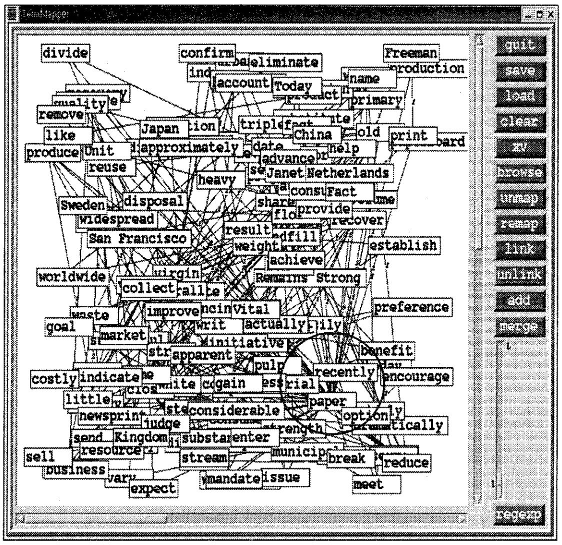

Figure 4 Sample for initial map（one document）

Figure 5 is the map which is redrew by cooccurrence frequency. It looks much better than figure 4. Frequency of cooccurrence is very important to look at hundreds of extracted key terms. When the document doesn't have enough sentences, all of the cooccurrence frequency will often become just one. Like this case, redrawing by cooccurrence frequency doesn't work. So users can choose the key terms from the system provided list which are drawn on the map and remap the relationships of key terms.

Figure 4 shows the result including the terms of which the frequency is just one. Table 3 shows the cooccurrence data which exclude the low frequency terms.

Table 3 The cooccurrence frequency and

cooccurrence key terms

6 recover & paper	2 paper & old
5 old & paper	2 paper & consume
4 recycle & paper	2 old & recycle
3 paper & recovery rate	2 recycle & program
2 writ & paper	2 Remains Strong & Remains Strong
2 virgin & paper	1 wastepaper & close
2 wood & pulp	1 waste & important
2 waste & paper	1 waste & disposal system
2 produce & paper	1 waste & disposal
2 paper & recycle	1 waste & call
2 waste & country	1 waste & Germany
2 solid & waste	1 pulp & paper
2 paper & recovery	1 paper & wastepaper
2 paper & produce	:
2 paper & paperboard	:

In figure 4 lower right, users can find with difficulty the main key term in the hundreds of words because the main key term links many other words. In this case, the key term is "paper". If users decrease the number of key terms by frequency, the key tem like "paper" stand out on the map. This is the advantage of visualizing information by frequency of cooccurrence on the map.

Figure five shows the key terms of which frequency of cooccurrence is more than twice. This is more understandable compared to former map because the only high frequency terms and their relationships are visualized.

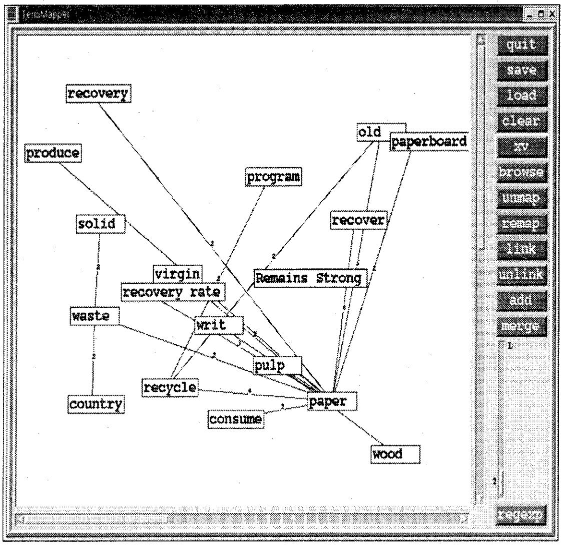

Figure 5 Sample of the map after decreasing the number of key terms by frequency

In the map of Figure 5, users can see the number on the line. It is the number of cooccurrence frequency. Users can rearrange the key terms on the map like figure 6 according to this number. Rearranged map by author is like figure 6. It is more organized and displays clear relationships of key terms. Users can figure out which key terms are related to the main key terms directly. The most important part of this system is to tell the key terms which are important but unexpected for users. In figure 6, the relationship of "paper--recycle--program" is the example. It tells users that programs for recycling paper are very important when users research about paper.

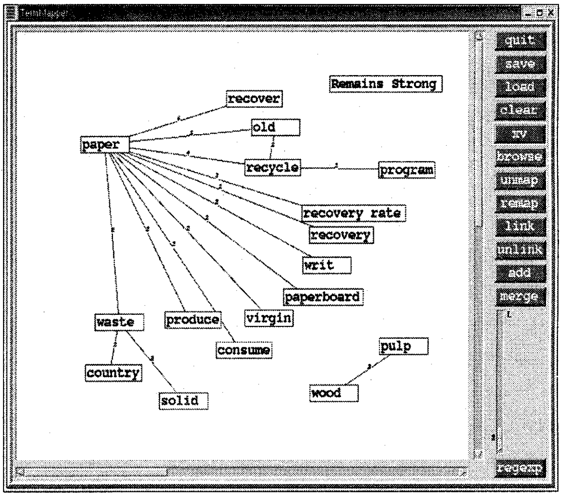

Figure 6 Rearranged map by author

6. Discussion

In figure 6, the terms "wood" and "pulp" seem to be separated from the main key terms "paper". But document in the database describes the relationships about paper and pulp. So these two key terms is linked each other on the same map by cooccurrence frequency just one. This happens because the document doesn't have enough cooccurrence frequency to display the relationships of paper and pulp. For this kind of problem, this system provides the function to go back to the documents and remap function for hided key terms.

Ending of Table 3, we can find the data "1 pulp & paper" and "1 paper & wastepaper". In Figure 7, wastepaper is selected from the list by users and then system draw link between paper and pulp. Also system draw links wastepaper and other indirect connected key terms. In this way, paper and pulp linked directory again

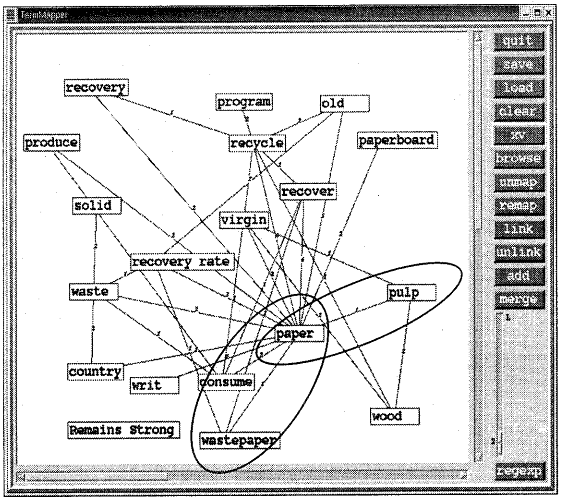

Figure 7 Remapping hided key terms and links

According to the test subjects, this system has advantages to understand the context or generate the idea [5]. This means the system succeed to visualize the relationships of key terms. For better map, it is necessary to install grammatical approach. The disadvantage of this system is that the system is not enough to substitute for reading the documents. The displaying way is another problem. It still depends on human skill to organize key terms. To develop the system to arrange the terms structurally is our next goal.

7. Summary

In this paper, we discussed the TermLinker system which is for visualizing information. The process to visualize the relationships of key terms is composed of three steps. The first is searching key terms (term, compound, high-frequency word) from the documents. The second is connecting the cooccurrence relationships. The third is visualizing the relationships on the map. The final step is the most important part in this study because it is the system that can display the knowledge chain automatically. It is necessary to develop the way to search the word not only by its frequency of cooccurrence but also by grammatical function.

Acknowledgements

We would like to acknowledge Aichi University Research Grant.

Reference

[1] Card, S. K., Mackinlay, J.D., Shneiderman, B.: “Readings in Information Visualization Using Vision to Think,” Morgan Kaufmann, pp.686 1999.

[2] Chang,Ch-H. Hsu, Ch-N, Lui, Sh-Ch, “ Automatic information extraction from semi-structured Web pages by pattern discovery,” Decision Support Systems, vol. 35, no. 1, pp. 129-147, Apr 2003.

[3] Chaomei, Chen: “ Information Visualisation and Virtual Environments, Springer Verlag, pp.223 1999.

[4] Fayyad. U. M., Grinstein G.G., Wierse A.: “Information Visualization in Data Mining and Knowledge Discovery,” Morgan Kaufmann, pp.407 2002.

[5] Hori, Koichi.: “Concept Space Connected to Knowledge Processing for Supporting Creative Design,” Knowledge-Based Systems Vol.10, No.1, pp.29-35 1997.

[6] Konomu, Dobashi., “Information visualization and problem finding, Arm, pp.290 Feb 2000 (Japanese).

[7] Leponiemi, J, “Visualizing discussion history,” International Journal of Human-Computer Interaction, vol. 15, no. 1, pp. 121-134, 2003.

[8] Perrin, P; Petry, F E, “Extraction and representation of contextual information for knowledge discovery in texts, Information Sciences, vol. 151, pp. 125-152, May 2003.

[9] Yang, Ch C, Chen, H, Hong, K, “Visualization of large category map for Internet browsing, Decision Support Systems, vol. 35, no. 1, pp. 89-102, Apr. 2003.