社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133 (31)
2003／9／16

Information Clipping from Internet Documents with Similar Contexts

Eiji MURAKAMI† and Takao TERANO†

†：Graduate School of Business Sciences, University of Tsukuba, 3-29-1, Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
E-mail: †murakami@gssm.otsuka.tsukuba.ac.jp, †terano@gssm.otsuka.tsukuba.ac.jp

Abstract There are so many documents available in the Internet. Some of them implicitly share common contexts. The examples of contexts covers pre-determined tasks, i.e., sales reports, categories, i.e., concept hierarchies, and forums, i.e., special interest groups. By clipping, we mean (1) to define the importance measures of documents in the same context, and (2) to acquire the important statement(s) from the documents based on the measure. This paper describes a new method of information clipping suitable for the group of documents gathered from a certain context retrieved in the Internet. The basic steps of the method is (1) to get key words using KeyGraph from a given set of documents, (2) to cluster the documents by applying Dulmage Mendelsohn decomposition algorithm for bipartite graphs, which consist of the nodes of the important words and the documents and the edges to represent their inclusion relationship, and (3) to acquire the corresponding important sentences. The paper shows some experimental results to reveal the effectiveness of the proposed method using a prototype system applied to the practical internet documents.

Keyword Information clipping, clustering, summarization

共通性のあるインターネット上の文書からの情報クリッピング

村上英治†　寺野隆雄†

†筑波大学大学院ビジネス科学研究科, 112-0012 東京都文京区大塚 3-29-1
E-mail: †murakami@gssm.otsuka.tsukuba.ac.jp , †terano@gssm.otsuka.tsukuba.ac.jp

あらまし 最近はフォーラムなどに代表されるようにある程度話題の共通性が保障されるような文章がインターネット上に多く存在するようになってきた。また、社会科学的な視点でこのようなインターネット上の文章を分析することでインターネットの中の社会、ひいては現実の社会のさまざまな現象を説明したり近い将来発生する可能性のある現象を予測したりする研究も行われている。２００２年４月には世界最大の検索サイトである Google は彼らの検索エンジンへの API を公開しインターネット上のコンテンツを誰でも簡単に入手できるようになった。本論文ではある程度共通性のある文章集合を入手したあとの課題として情報量を減らしながら重要な内容だけを残す情報クリッピングが重要であると考え、そのための手法を提案する。

キーワード 情報クリッピング、クラスタリング、要約

1. Introduction

Recently, we are able to find sufficiently good information from the messy Internet world, if we could fully utilize conventional search engines such as Google. Moreover, recently experiments has started, which use search engines not only by human users from a Web browser but by computer programs. For example, through Google Web API [1], we can obtain the information over 2 billion Web pages stored in Google.

Understanding the corrent affairs, in this paper, we will focus on the information clipping method applied to the documents downloaded into users' computers from the vast amount of the Google's world. By information clipping, we mean (1) to define the importance measures of documents in the same context, and (2) to acquire the important statement(s) from the documents based on the measure. We assume such clipped documents share a common context. We propose a new information clipping method to get only important contents with small amount but rich information.

The contents of the paper is summarized as follows: in Section 2, we discuss why and how to get the documents with the same context from the Internet. In Section 3, we will describe the proposed information clipping method in detail. In section 4, we show some experimental results and in Section 5, concluding remarks are given.

2. Why and How to Get Documents from the Internet

Every day, people get brought information using the web browser for some purpose. The activities include, for example:

- Surveillance of specific information.

They check whether the new information has been acquired from the results via conventional search engines with fixed keyword sets. Then, the new information is used to re-examine the contents.

- Looking for some solutions

They perform netsurfing using search engines to find some good contents, which will derive some solutions for the specific problems.

In this paper, we assume that we know where the URLs of useful contents are. The information clipping tasks will start just after these contents are downloaded in to their own computers using conventional search engines.

Google is one of such conventional search engines. In April, 2002, they opened Google Web API, which allow users to use Google from their own computer programs. Google Web API is the technology for Web services based on SOAP 1.1 (Simple Object Access Protocol) and WSDL (Web Services Description Language). Google Web API will supports the basic functions of searching , caching , spell checking and the correcting used in user programs.

Fig. 1 shows an example screen to search for the word 'tokye' from the web browser of Google. After completing the process, Google has presented the result of the link information to Web contents. Gooogle automatically change the word 'tokye' to 'tokyo' using the spell checking and correcting functions.

Google prepares Google Web APIs Developer's Kit which includes a WSDL file to define the sample of the Java client programs using the SOAP interface. So Google Web API can be easily used by users referring this WSDL file.

If URLs or search key words are already known, the contents are easily retrieved by the conventional technology. That is, Google Web API will be used to get some set of Web contents on users' own computers. After the contents with the same contexts are downloaded into the users' computers, then the information clipping tasks will start to process the contents resided on their own computers.

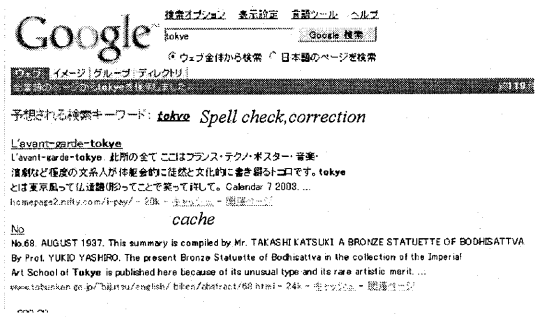

Fig.1 Example result of search key word of tokye in Google

3. Information Clipping Tasks

By clipping, we mean (1) to define the importance measures of documents in the same context, and (2) to acquire the important statement(s) from the documents based on the measure. This paper describes a new method of information clipping suitable for the group of documents gathered from a certain context retrieved in the Internet. The basic steps of the method is (1) to get key words using KeyGraph from a given set of documents, (2) to cluster the documents by applying Dalmage Mendelson decomposition algorithm for bipartite graphs, which consist of the nodes of the important words and the documents and the edges to represent their inclusion relationship, and (3) to acquire the corresponding important sentences.

We explain the detailed methods mentioned above in following sub-sections.

3.1. Find the Key Words

We assume that each document is constituted by two or more sentences with the following special feature. Some sentences in a document have the lexical chain [2] for rest of other sentences. This is the consistent concept, that is, the word in each sentence is mutually related to each other if the lexical chain exists. Such words have a role to control the flow of context in the document. To identify such words, the algorithm of KeyGraph [3] is adequate. KeyGraph focuses on the features of the lexical chain being used, and extracts the corresponding keywords in a set of statements. Although KeyGraph was originally

proposed for the purpose of extracting the keywords of the whole statements. We apply KeyGraph to the key word extraction from a given set of documents (one document consists of plural sentences), which share the same contexts exhibited by the lexical chain. We suppose that a word with the following feature would be a candidate of a keyword contained in plural documents in the same context.

First, we set a word set W with high frequency of appearance from the whole documents:

W= [w1, w2, --, wn]. (1)

And #(wi) represents the frequency of the word wi in the documents.

Then, We determine keywords based on the importance index, which is calculated from the frequency of co-occurrence of all two word combination in each document. That is, when a co-occurrence relationship exists in the pair of the word in W, the importance measure Pi is calculated as follows.

Cij $(i \neq j)$ is an index to specify the co-occurrence:

Cij = 1 when wi co-occurs with wj.

Cij = 0 when wi does not co-occurs with wj.

$$Pi = \frac{\#(wi)}{n} \times \sum_{j} \frac{\#(wj)}{n} Cij \qquad (2)$$

If Pk is positive for any words wk, then we define a set of such wk is a key word set K.

K= [k1, k2, --, km]. (3)

3.2. Document Clustering via DM Algorithm

Let pl be a pair of keywords (ki, kj), where ki, kj ∈ K and let di be a document in the document set. N is the number of the documents. Using the notation, we define

matrix $A(N, m^2)$ (4)

where Aij = 1, if the pair pi =(ki, kj) is contained in the document dj and Aij=0 otherwise.

Where if co-occurrence pair word pi is involved to dk, value of pi is 1. If it doesn't, value of pi is 0. Therefore we obtain the matrix A which is showing the relationship between each document and co-occurrence pair word that is only expressed as 0 or 1.

$$A = \begin{array}{c|ccc|} & & & \\ p1 & & & \\ & & & \\ & & & \\ pm^2 & & & \\ \hline & d1 & & dn \end{array} \qquad (5)$$

When replacing each row and for each column, we can obtain canonical form A' of the matrix A, which is attained by Dulmage Mendelsohn decomposition [4][5].

$$A' = \begin{array}{|ccc|} A00 & & \\ & A11 & * \\ & & \\ 0 & Arr & \\ & & Ar+1,r+1 \end{array} \qquad (6)$$

The feature of A' are

(i) A lower left entry of A' which is below the A00,...,Ar+1,r+1 that are diagonal block of A'. All those entry are zero.

(ii) A11,...,Ar,r are a sub square matrix.

The matrix A and its canonical form A' would be the changes of node names in bipartite graph G.

G=G (T, D;E) |T|= m^2 and |D|=n. (7)

Where we define a vertex set that are T= {p1, ..., pm^2 } and D= {d1,..., dn} and follows.

$$\mu ij (i = 1,...,m^2; j = 1,...,n)$$

$$\mu ij = 1, if (pi, di) \in E$$

$$\mu ij = 0, if (pi, di) \notin E$$

At this time μij is i,jth entry of m x n matrix. So we

have M(G) =[μij], this is bipartite graph G. Conversely if M (G) is 0-1 matrix, we can give 0-1 value to the arbitrary $m^2 \times n$ matrix of A=[μij]. This would be bipartite graph G. If A is graph matrix, we can obtain A' which is canonical form of A. The feature of A' is below.

(iii) A00, ..., Ar+1,r+1 are irreducible diagonal blocks of G (A'). Each block can be consider as a graph matrix of G0,...,Gr+1. Especially A11,..., Arr are a sub square matrix.

We find Gii that is bipartite graph matrix from real document set then draws actual bipartite graph shown in Fig.2. As shown in Fig. 2, if vertex of p defined by a formula (7) and vertex of d that represents document, we define the some of the documents are clustered by the some of the p.

3.3. Extraction of Key Sentences

In the formula (6) the diagonal line of every sub square matrix aligns 1 that calls core of matrix A'. Fig. 3 is the example of A' which can be seen matrix core is made from actual document clustering. Moreover, this core equals to the maximum matching of bipartite graph of a formula (7).

The each document d is consists of sentences si.

d= {s1, --, si, --} (8)

Therefore how to extract the key sentence s from document d is

(a)Find pi which is the element of the feature vector defined by the formula (7) and is the matching partner of each document di because these are the vertex of maximum matching of bipartite graph.

(b) The sentence s is consists of some words and the document d is consists of some sentences. Some sentences s containing pi= {ki, kj} can be extracted for each document.

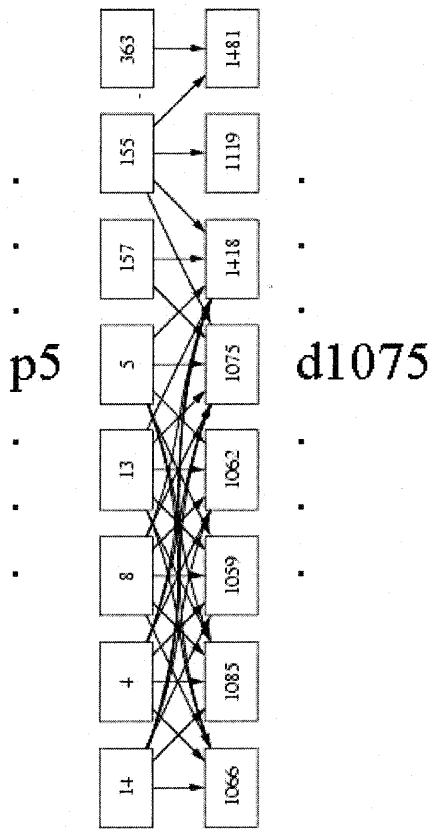

Fig. 2 Example of Gii obtained from the actual document set

4. Experiment result

We prepare a test data shown in Fig. 4. Each raw of the data consists of one document with the following attributes: document ID, document Description, attribute1, attribute2, attribute3 and Result.

ID, attribute1, and attribute3 are numerical attributes, and attribute2 and Result are nominal attributes. Description is a text attribute. We have used the collection of 145 documents then applied the proposed method to the data. At this time, we only used Description attribute as a target attribute.

The graph matrix obtained from the sample data for this experiment results in the clustering shown in Fig. 5. The diagonal entry of this matrix from the upper left to lower right looks like line that is Aii entry defined by a formula (6). Some portion of the thick line with a wen like form represents sub square matrix.

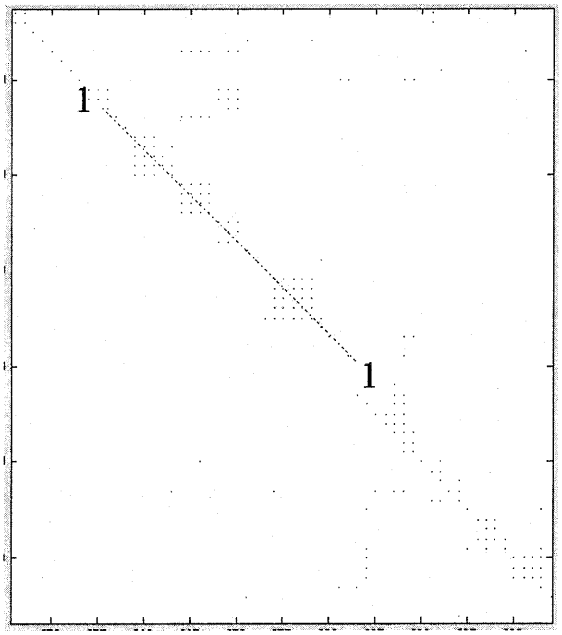

Fig. 3 Example of A' by the actual document clustering

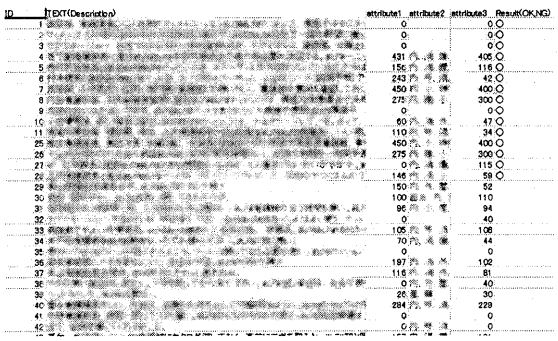

Fig.4 Example of sample data

Fig.6 shows recombination of the original documents represented in Fig.5. Fig.6 means the relationship between pi= {ki, kj} and each document d along with the lower right to the upper left of core of matrix in Fig.5. This is the maximum matching of bipartite graph in Fig. 5. From Fig.6, we observe that

(a) Let ID, pi= {ki, kj}, Description, Result, and ClusterID is row in table.

(b) ClusterID is given to the each Description associated with the core of a matrix. Since the core of matrix exists on the diagonal line of a matrix, same ClusetrID could be assigned if the core of matrix belongs to the same sub square matrix.

(c) Yellow cells represent the clustering results in which multiple document belongs to the same sub square matrices.

(d) The result of document clustering resulted in the classification of the documents into the two classes: Result(OK) or Result(NG).

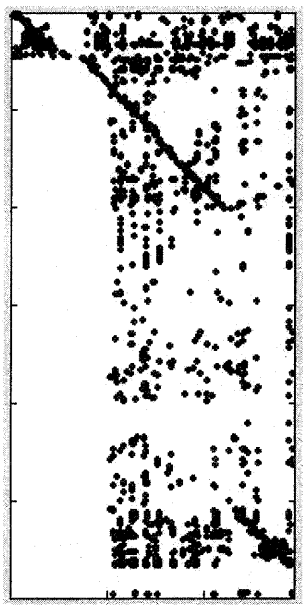

Fig.5 The bipartite graph matrix from sample data.

Fig.6 The clustering result of sample data.

Next, we choose 28 documents, whose values of Result are only Result (O.K.) from 145 documents, then applies the proposed method to the Description attribute. The result is shown in Fig. 7. We have a comparatively big sub square matrix in the middle of a graph matrix as a clustering result. This represents the relationship between Result and Description with Result (O.K.) is discovered. We can understand the context of typical Description contents (Result is Result (O.K.)).

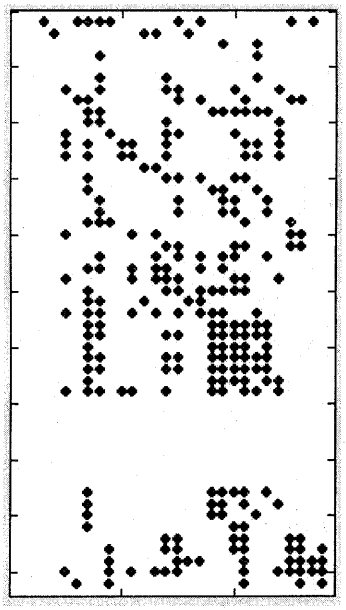

Fig.7 The bipartite graph matrix when Result (O.K.)

The extracted key sentences in the document are shown in Fig. 8. The key sentences can be extracted from documents, which contain pi= {ki, kj}..

Fig.8 The example of extraction of the key sentence using sample data

In Fig.8 one topic sentence is extracted from the three sentences the Description..

5. Conclusion

This has proposed a method for information clipping from internet documents with a similar context. To get the documents in a similar context, it is beneficial to use conventional search engines such as Google and corresponding APIs, however, to extract important information, the proposed method will be of use.

The information clipping method will contribute to

- Discovers some meanings of the target context from the gathered documents by clustering them;
- Extract key sentences from the clustered document.

We have already found interesting hypotheses in the other large document sets using the proposed method, although we have not reported the results in the paper, which will be presented elsewhere.

6. Acknowledgement

The research is supported in part by Grant-in-Aid for Scientific Research of Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan: Priority Area, Active Mining (13131202), and Informatics (15017206). We also expresses our thanks to Prof. Yukio Osawa, University of Tsukuba, for his essential and suitable advice for this research.

Reference

[1] Google Web API, http://groups.google.com/apis/

[2] J. Morris and G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," Computational Linguistics, vol.17, no.1, pp.21-48, 1991

[3] Yukio Osawa and Nels E Benson and Masahiko Yachida, "KeyGraph: Automatic Indexing by Segmenting and Unifing Co-occurrence Graph," IEICE D-I Vol. J82-D-I No.2 pp.391-400, 1999

[4] A. L. Dulmage and N.S.Mendelsohn, "Covering of bipartite graph," Canad Jour. Math. , 10, pp.517-534 (1958).

[5] A. L. Dulmage and N.S.Mendelsohn, "A structure theory of bipartite graphs of finite exterior dimension," Trans Roy. Soc. Canad. , Sec. III, 53, pp.1-18 (1959).