Graph-based Text Database for Knowledge Discovery

Junji Tomita
NTT Cyber Space Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239-0847 Japan
tomita.junji@lab.ntt.co.jp

Hidekazu Nakawatase
NTT Cyber Space Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239-0847 Japan
nakawatase.hidekazu@lab.ntt.co.jp

Megumi Ishii
NTT Cyber Space Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239-0847 Japan
ishii.megumi@lab.ntt.co.jp

ABSTRACT

While we expect to discover knowledge in the texts available on the Web, such discovery usually requires many complex analysis steps, most of which require different text handling operations such as similar text search or text clustering. Drawing an analogy from the relational data model, we propose a text representation model that simplifies the steps. The model represents texts in a formal manner, Subject Graphs, described herein, provides text handling operations whose inputs and outputs are identical in form, i.e. a set of subject graphs. We develop a graph-based text database, which is based on the model, and an interactive knowledge discovery system. Trials of the system show that it allows the user to interactively and intuitively discover knowledge in Web pages by combining text handling operations defined on subject graphs in various orders.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models, search process*

General Terms
Algorithms, Design

Keywords
Subject Graphs, Knowledge Discovery, Interactive Search

1. INTRODUCTION

The Web is a huge, diverse and dynamic information source, and knowledge can be discovered in the texts available on it. Knowledge is defined here as important facts that support a person in making a decision. Knowledge discovery usually requires complex analysis steps that consist of a heterogeneous combination of text handling operations such as searching, clustering, summarizing, and comparing[3]. As for structured data, Relational Data Model(RDM) enables several data handling operations to be combined, because it represents data in a formal manner, i.e. relations, and provides data handling operations whose inputs and outputs are identical in form, i.e. a set of relations[1].

By analogy with RDM, we propose a graph-based text representation model that represents texts in a formal manner, i.e. Subject Graphs[6] and provides text handling operations whose inputs and outputs are identical in form, i.e. a set of subject graphs. Subject Graphs are an extension of term vectors and are made automatically by calculating the significance of term-term associations in addition to that of terms. Each node and link has a weight corresponding to the significance of term or term-term association, respectively. The model can separate text data from the application programs, and enables various text handling operations to be flexibly combined. We have developed a graph-based text database (GTB) based on the model and an interactive knowledge discovery system that uses GTB. Trials of the system show our model can be effective for knowledge discovery.

Copyright is held by the author/owner(s).
WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

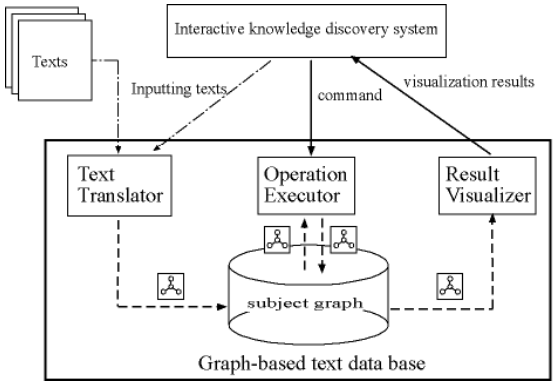

Figure 1: Graph-based text database and interactive knowledge discovery system

2. GRAPH-BASED TEXT DATABASE

Figure 1 overviews the GTB and an associated interactive knowledge discovery system. **The text translator** translates each text into a subject graph by 1)extracting terms from the text, 2) calculating the significance of each term from its occurrence frequency using a variant of the BM25 method[4], and 3)calculating the significance of each term-term association from the co-occurrence frequency of two terms in a unit such as sentence, clause, or a word window. **The operation executor** executes graph handling operations whose inputs and outputs are a set of subject graphs. To achieve a full set of analysis steps, we define 6 operations: (a)searching for similar graphs, (b) clustering graphs, (c) extracting partial graphs, (d) adding graphs (merging

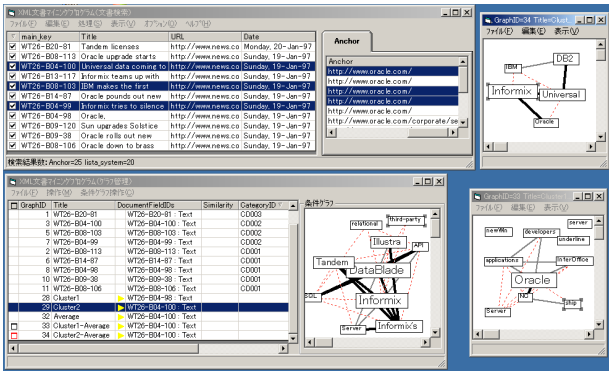

Figure 2: The user interface of the interactive knowledge discovery system

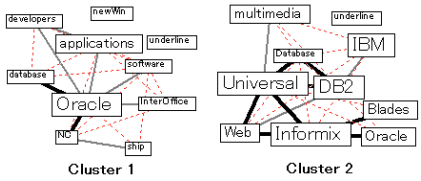

Figure 3: Comparison of two clusters

and averaging graphs), (e)subtracting graphs, (f) selecting graphs by attributes. The details of these operations are described in [6] and the next section. These operations can be combined in any order, because all of them input and output a set of subject graphs. Although we define here only 6 operations, any operation whose inputs and outputs are a set of subject graphs can be incorporated into the module. **The result visualizer** visualizes the operation results by using the spring model[2]. Users can understand the significant terms and the term-term associations from the visualization results, and so can grasp the contents of the found texts.

3. TRIALS

Figure 2 shows the user interface of the system. The user can invoke text handling operations in different orders interactively and intuitively through these windows. 247,489 Web pages in the WT2G test collection[1] were used as the targeted text data. The goal of the user is to discover knowledge such as 'Who is the best relational database vendor?', How do users rate the products?' and 'Are there any competitive products?'. Using the system, the most probable analysis steps are as follows.

The user inputs the expression 'relational database' and invokes (a). The system locates 253 Web pages similar to it. The user invokes (f) with the condition specifying an anchor attribute contains the name of database vendor. The user discovers that the Oracle site has the most references, 11. The user invokes (d) to make a merged graph (g_o) to overview of all the pages referencing the Oracle site, and (b) to cluster them with similar contents. The system classifies them into 5 categories (7,3,1,1,1). The user invokes (d) for the pages in each category, (e) to subtract g_o from each merged graph, and (c) using 'relational database' as the condition. The system merges graphs in each category, subtracts g_o from each merged graph, and extracts the partial graph that has nodes neighboring to 'relational database' from each subtracted graph. These operations compare the contents of each category by focusing on 'relational database'. Figure 3 shows two graphs from Cluster 1 and Cluster 2. From Figure 3, Cluster 1 seems to be relevant to InterOffice, while Cluster 2 seems to compare Oracle with Infomix or DB2[2].

In brief, from these interactive steps, the user can discover knowledge such as (A)The relational database vendor that is attracting attention is Oracle. (B)The name of Oracle's product is InterOffice. The rating and specification of the product may be found in the pages in Cluster 1. (C) Its competitors are Informix and DB2. Comparisons may be found in the pages of Cluster 2. In this way, the interactive steps offered by the system reduce the user's effort in discovering knowledge in Web pages.

4. RELATED WORKS

Term vectors are widely used for formal text representation. Compared to term vectors, Subject Graphs allow the similarity between texts to be calculated more precisely and provide better visualization results by incorporating term-term associations. Furthermore, Subject Graphs support complex operations such as extracting partial graphs, which is not possible if only term vectors are used. Conceptual Graphs[5] can be used to represent texts. Although they may represent the content of a text in a sophisticated way, applying them to Web pages is difficult. This is because they are based on deep analysis, and so require well maintained dictionaries and an excessive amount of time to operate.

5. CONCLUSION

We proposed a text representation model that allows a wide variety of text handling operations to be combined for realizing the complex analysis steps needed to discover knowledge. We implemented the model by using Subject Graphs as the formal text representation. A graph-based text database based on the model and an interactive knowledge discovery system were implemented. Trials confirmed that the proposed model reduces the user's effort and is effective for discovering knowledge in texts on the Web.

6. REFERENCES

[1] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[2] P. Eades. A heuristic for graph drawing. *Congressus Numerantium*, 42:146–160, 1984.

[3] M. Hearst. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics(ACL'99)*, pages 3–10, 1999.

[4] S. E. Robertoson and S. Walker. Okapi/Keenbow at TREC-8. In *NIST Special Publication 500-246: The Eighth Text REtrieval Conference(TREC8)*, 1999.

[5] J. F. Sowa. Conceptual graphs for a database interface. *IBM Journal of Research and Development*, 20(4):336–357, 1976.

[6] J. Tomita and G. Kikui. Interactive Web search by graphical query refinement. In *Poster Proceedings of the 10th international World Wide Web conference(WWW10)*, pages 190–191, 2001.

[1] http://trec.nist.gov/

[2] InterOffice is an Oracle product and Informix is from IBM. DB2 is IBM database software.