A system that supports discovery of unexpected information from Web-documents

Daisuke Noguchi (710056)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 5, 2009

Keywords: Web search, Data mining, Graphical user interface, Hyponymy relations, Semantic similarities.

In this thesis, we propose a method for supporting keyword search to obtain "unexpected information" based on the analysis of HTML documents in the Web. More precisely, we propose a Graphical User Interface (GUI) that enables users to easily find unexpected but useful information and a scoring method for rating the unexpectedness of a combination of two search keywords. The GUI was already implemented in our search support system named "TORISHIKI".

In the high-speed development of the recent information-oriented society, it has become a common practice in daily lives to find useful information in the Internet using search engines. The problem is that it is often quite difficult to find appropriate search keywords to obtain proper information. In order to help users find such keywords easily, our group developed a search directory named "TORISHIKI". The directory can provide many keywords related to a given query and related keywords are classified into semantic categories. We assume that users typically submit a query to search engines with the intention of "using" the object referred to by the query and that information concerning troubles, ideas and tips in the context of the use of the object are useful for users. Based on this assumption, we prepared "troubles", "methods/tips" and "tools/materials" in the use context as semantic categories of terms related to the given query. As an

example; assuming a user gives a query "DHA", which refers to a health food supplement. Then, TORISHIKI provides "bleeding" as one of the troubles relating to DHA. Actually, it has been reported that excessive intake of DHA can interfere with the body's ability to stop bleeding when injured. This kind of information is unexpected or unknown to many but nonetheless quite useful for users who intend to use "DHA". Our aim is to help users find such unexpected but useful information easily. The related keywords are obtained from a large collection of Web documents and automatically classified into semantic categories. In this thesis we describe the GUI through which users can look for this kind of unexpected but useful information. Our GUI has more specific features as follows.

1. Terms related to a given query are displayed in a two-dimensional plane, and they are layed out according to the semantic similarities among them and the relation strength which is estimated using co-occurrence frequencies of the related terms and a given query.

2. Query expansion and generalization using the query terms' hyponymy relations.

As for No.1, the related terms are displayed in positions surrounding a category name such as troubles, which is located in a certain fixed position. The distance between a related term and the category name is computed based on co-occurrence frequencies. Roughly, the larger the co-occurrence frequency gets, the smaller the distance becomes and the related term is located closer to the category's center. On the other hand, the distance between related terms is determined automatically by semantic similarities computed by an EM-based clustering method. The more similar two terms are, the closer they are positioned. As for No.2, we developed a query generalization method based on hyponymy relations that were automatically acquired from the Web. TORISHIKI can display terms related to a hypernym of a given query by using this mechanism. We also conducted experiments in which we asked 60 graduate students or researchers if it is possible to use TORISHIKI for finding useful keywords not found in commercial search engines. 70% of the users answered "yes" to this inquiry.

We also developed a scoring method that rates the unexpectedness of combinations of two search keywords. We have developed the following four scoring schemes.

1. The co-occurrence frequency between a topic and a related term.
2. Mutual information between a topic and a related term.
3. The scoring scheme based on the distribution of related terms among the words similar to a query.
4. Frequencies of related terms.

Topic	Trouble expression	Comment
Cotton	Pesticide damage	The adhesion of the defoliant to the cotton
Cotton	Dust mite	Cause of house dust in a ball of dust
Historical play	Overproduction	A dispute of Kokuryo between China and Korea
Trademark rights	Dilution of equity	The generalization of the trademark
Trademark rights	Extinction	The extinction of the right by non-update
Brain abscess	Infectiousness disease	Bacteria arriving in the brain through the blood cause pus to pile up
Japan successive cabinets	Crime	About the peace treaty "Constitution of Japan"
Yin Yang fortune-teller	Assassination	An assassin disguised as a Yin Yang fortune-teller
Factory wastes	Sludge	A technical term concerning liquid waste management
Carbon	Stick	"The black dirt" which is seen in a ship
Forest	Desiccation	A factor of the salamander decrease
Sunflower	Gray mold disease	A technical term
Sunflower	Purple blotch	A technical term
Bhutan	Child trafficing	child trafficing in Indian border neighborhood
Bee	Feces pollution	The pollution at the apiary outskirts

Table 1: Examples of unexpected information, with comments

In a series of experiments we show that the last scheme shown above works best. Some samples of unexpected information derived from our experiments is shown in Table 1.