Building Term Suggestion Relational Graphs from Collective Intelligence

Jyh-Ren Shieh[1] Yung-Huan Hsieh[1] Yang-Ting Yeh[1] Tse Chung Su[1] Ching-Yung Lin[2] Ja-Ling Wu[1]

[1]Dept. of Computer Science and Information Engineering
National Taiwan University
Taipei 106, Taiwan
{jerry, ejection, dy, jsrf, wjl}@cmlab.csie.ntu.edu.tw

[2]IBM T. J. Watson Research Center
Hawthorne, NY 10532,
USA
chingyung@us.ibm.com

ABSTRACT

This paper proposes an effective approach to provide relevant search terms for conceptual Web search. 'Semantic Term Suggestion' function has been included so that users can find the most appropriate query term to what they really need. Conventional approaches for term suggestion involve extracting frequently occurring key terms from retrieved documents. They must deal with term extraction difficulties and interference from irrelevant documents. In this paper, we propose a semantic term suggestion function called Collective Intelligence based Term Suggestion (CITS). CITS provides a novel social-network based framework for relevant terms suggestion with a semantic graph of the search term without limiting to the specific query term. A visualization of semantic graph is presented to the users to help browsing search results from related terms in the semantic graph. The search results are ranked each time according to their relevance to the related terms in the entire query session. Comparing to two popular commercial search engines, a user study of 18 users on 50 search terms showed better user satisfactions and indicated the potential usefulness of proposed method in real-world search applications.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]; H.3.3[**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering*, *Selection process*; I.7.1 [**Document and Text Processing**]

General Terms

Algorithms, Human Factors, Experimentation

Keywords

Social Network, Keyword Expansion, Re-ranking.

1. INTRODUCTION

We propose to utilize the collaborative knowledge of human beings stored on Wikipedia[1] to generate semantic graphs of suggested search terms. A semantic graph of a given search term is centered on the term, and there are tons of related search terms that link to each other with various weights on the links. It is similar to an ego-centric social network but uses terms to replace people in nodes. We propose to use an important hidden network, which links Wikipedia's contributors (i.e., article editors) to key terms on Wikipedia pages, to consider the relatedness of terms as well as how close the terms are. In our work, the calculation of the relatedness of two terms does not base on whether the terms are hyperlinked through the "related links" in a Wikipedia page, because this approach cannot weight the semantic proximity of terms. Our hypothesis is that when a single editor contributes to two articles in which the two terms occurred separately, there is certain likelihood that these two terms are somehow related based on high clustering coefficients of social networks. We thus draw a complex semantic graph between contributors and terms and conduct to one-dimensional graphs with weighted links. We propose a Collective Intelligent based Term Suggestion (CITS) system, which provides novel semantic related term suggestions as semantic graphs. The rest of this paper is organized as follows. In Section 2, we present our framework of semantic graph based term suggestion functions. User evaluation results of the proposed system are reported in Section 3. Finally, conclusions and future work directions are provided in Section 4.

WWW 2009, April 20-24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. A FRAMEWORK FOR PROVIDING SEMANTIC TERM SUGGESTION

Each Wikipedia document contains related terms which form a kind of conceptual network. For each article, all versions and corresponding contributors are stored. Social networks of contributors and their edited articles can then be formed. These two networks provide basic elements for computing semantic relatedness. We crawled thousands of Wikipedia articles to obtain hyperlinks and contributors' information from editorial histories for constructing the relationships between terms. Beginning from each topic page, we crawled all of the topic's internal links to create a concept network. In addition, for each article a record of all contributors and their editorial histories is maintained in Wikipedia. We identified the relationships between these articles and their respective contributions as a new type of network, a specific bipartite network.

Let us consider a topic "Tm = Wii" and its related articles as an example, we treat "Wii" not just as a topic but also as a key term which we denote as T_m^0 . Beside "Wii", there are many internal hyperlinks linked to Wii-related terms, within the articles, focused on different topics such as "Nintendo" and "Xbox 360," which are denoted as T_m^{Hi}. So, a set of Wii-related terms can be created and is denoted as $T_m = \{T_m^0, T_m^{H_1}, T_m^{H_2}, \cdots, T_m^{H_n}\}$. We identify the relationship network among internal hyperlinks as the concept network.

For each article, we examine the latest 500 revisions of article's editorial history $RT_m = \{RT_m^{H_0}, RT_m^{H_1}, RT_m^{H_2}, \cdots, RT_m^{H_n}\}$ of T_m , where RT_m^H denotes the set of records corresponding to $T_m^{H_i}$. On the basis of this information provided by Wikipedia, we designed a parser to ignore anonymous contributor so we can see all of the real contributors who have contributed to those specific set of topics. After this procedure, we can create lists of

contributors, $CT_m = \{CT_m^0, CT_m^{H_1}, CT_m^{H_2}, \cdots, CT_m^{H_n}\}$, who have contributed to any of the hyperlinked terms in Tm.

We group together all of the terms contributed by a group of contributors as an effective sampling. Indeed, the number of all of the terms contributed by this group of contributors CTm is definitely larger than the number of specific related terms Tm. Each contributor may have his diverse interests. Thus, this grouping procedure reflects the semantic relatedness buried in human behavior. This step introduces a useful sampling scheme for semantic relatedness and paves a way to locate important information inside the human centric concept and thus extend the relatedness of each term.

The topic-contributor bipartite graph can be represented as $G_{TC} \equiv \langle V_T + V_c, E_{TC} \rangle$, where $V_T = T_1 = \{t_i\}$ and $V_C = C_1 = \{c_i\}$denote the vertices of topics and contributors, respectively. For a topic-contributor bipartite graph consisting of n topics and m contributors, we can express the editing relationship by a binary matrix$B_{TC} = [b_{ij}]_{m\times n}$, where the element

$$b_{ij}=\begin{cases}1,\text{if the } i_{th}\text{conributor has edited the } j_{th}\text{topic,}\\ 0,\text{else.}\end{cases} \quad (1)$$

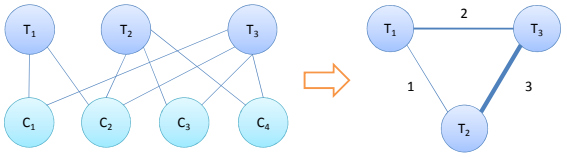
More specifically, the i_{th} row vector, $b_i \equiv \{b_{i_1}, b_{i_2}, \cdots, b_{i_n}\}$, of BTC denotes whether the topics have been edited by the i_{th}contributor or not. We then fold this bipartite graph into a topic only graph, which is called the semantic graph, by identifying the number of common members between contributor vectors b_i, b_j via the following formula

$$W_{ij} = \sum_{k=1}^{\mu} b_{ik} \times b_{jk} = b_j \cdot b_j^T, \quad (2)$$

that is, the value of W_{ij} is computed by the inner product of b_i and b_j. This measurement can be generalized to the whole matrix B_{TC}. So the weights of the corresponding semantic graph can be expressed by

$$W_{m\times n} = B_{TC}B_{TC}^T \quad (3)$$

Fig. 1 illustrates our approach to compute the weight of semantic relatedness.



$$B_{m\times n} = \begin{bmatrix}1 & 1 & 0 & 0\\ 0 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\end{bmatrix} \qquad W_{m\times m} = B_{m\times n}B_{m\times n}{}^T = \begin{bmatrix}2 & 1 & 2\\ 1 & 3 & 3\\ 2 & 3 & 4\end{bmatrix}$$

Fig. 1. Computing semantic relatedness weight via inner product of vectors.

3. EVALUATIONS OF CITS SYSTEM

Table 1 shows a comparison example of the related terms constructed from the proposed method and from the WordNet [2]. We can see that the proposed method is arguably closer to most users' perception today. We take "Number Theory" as a testing example and compare with Google [3]. When keying in "number theory" into CITS, the words like "Chinese remainder theory", "Fermat's little theory", "Prime number theory", "Riemann zeta function", and "Goldbach's conjecture", were listed in the suggestion box. But, among the suggested words providing by Google, we can only find words "Number theory web" that is conceptually relevant.

Table I: A comparison of suggestion terms based on Wikipedia and WordNet for the "search engine".

Using Wikipedia	Using WordNet
Web search engine	Search engine
Google	Program
Google search	Programme
Search engine (computing)	Computer programee
List of search engines	Computer program
Microsoft	Software
Live Search	Computer software
Yahoo	Software package
AOL	Software system
Ask.com	Software program
Ebay	Package
MSN search	Thought
Massachusetts Institute of Technology	System
	Promulgation
Web Portal	Create by mental act
File Transfer Protocol	System of rules
Dot-com bubble	Info
Semantic web	Show
PageRank	Information

Eighteen students were invited to test 50 search terms and compared with the search recommendations with Google and Yahoo! [4]. Examples of the query terms included "number theory", "solar energy", "encryption", and "complex network", etc. The experiment showed that 21% users think CITS is much better, 47% better, 10% the same, 17% worse, and 5% much worse, as compared with Google. Yahoo! has provided two kinds of term suggestions. One of them makes more like a word-completion suggestion and the other, called "Explore concepts", also a conceptual term suggestion mechanism. User studies on the basis of the same group of participants showed for word-completion approach: 12% users think CITS is much better, 41% better, 15% the same, 27% worse, and 5% much worse, as compared with Yahoo!. Compare with the Yahoo!'s Explore Concepts approach, users studies showed 8% think CITS is much better, 39% better, 35% the same, 15% worse, and 3% much worse. A demo site is at http://www.cmlab.csie.ntu.edu.tw/CITS/.

4. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a new framework for performing human-centric searches with complex semantic relatedness. We see a potential advantage in using such semantic graphs for various search tasks in people's daily search activities. Compared with WordNet, our approach leverages knowledge bases that are orders of magnitude larger and more comprehensive. Compared with commercial search engines such as Google and Yahoo!, CITS results in better semantic related search suggestions, and is thus more activated due to its usage of real time Wikipedia-based social network structures. Our future work will focus on enhancing CITS's search capabilities by adding personalization and trend prediction to the design.

5. References

[1] Wikipedia, http://en.wikipedia.org/.

[2] C. Fellbaum, editor, "WordNet: An Electronic Lexical Database," The MIT Press, Cambridge, MA, 1998.

[3] Google, http://www.google.com/.

[4] Yahoo!, http://www.yahoo.com/.