IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.

ITP4546

Indexing by Conditional Association Semantics

Xiaowei Yan, Chengqi Zhang, Shichao Zhang, and Zhenxing Qin
Faculty of Information Technology
University of Technology, Sydney
PO Box 123, Broadway, Sydney NSW 2007, Australia
Tel: +61-2-9514 4534, Fax: +61-2-9514 4535
{xyan,chengqi,zhangsc}@it.uts.edu.au

ABSTRACT

Prevailing information retrieval methods are based on either term similarity or latent semantics. Terms are considered independently. This paper presents a new strategy for information retrieval, i.e., indexing by conditional association semantics. *In our approach, the conditional association semantics of terms will be considered during semantics indexing.*

1. INTRODUCTION

Data on the WWW are usually structureless, dynamical, undisciplined, uncertain, and enormous. A large number of information sources, with their different levels of accessibility, reliability and associated costs, present us with a complex problem of information gathering. On the other hand, search engines often return many thousands, even millions of results in response to a user query. It would be difficult for a user to browse so much information searched. In particular, it is an important challenge to identify which pieces of the information are really useful to the user. Therefore, there have been many intelligence-based methods for information gathering (or information filtering) from the WWW proposed in recent literature [3-6].

To reduce irrelevant information searched, this paper presents a new strategy for information retrieval, named as indexing by conditional association semantics. Conditional association semantics is a relationship among terms of a document and a query. We begin with giving the problem statement and some related work in Section 2. Then a synthesizing model by weighting is presented in Section 3. In Section 4, a relative synthesizing model for association rules from unknown data sources is described. In Section 5, we conclude this paper.

2. PROBLEM STATEMENT

Generally, a user query can be described by using natural language, keywords, or a database query language [5]. The simplest form of a user's query is a list of one or more keywords. Experienced users may state their queries in an appropriate form to get what they want. However, there are still many inexperienced users. A typical user does not have the aptitude of using Boolean logic statements. The user is not often an expert in the area that is being searched. He may lack the domain-specific vocabulary, and usually start searching with a general concept of the information required.

A limited knowledge of both the specific vocabulary in a particular area and what is exactly needed leads to the uses of inaccurate and misleading search terms. Even when the user is an expert in the area, the ability to select the proper search terms is constrained by lack of knowledge of the author's vocabulary. Each writer has his own vocabulary formed by his life experiences, environment where he grew up, and ability to express himself. Thus, an information retrieval system should provide tools to overcome the search specification problems discussed above, and automatically assist a user for developing a search specification that represents both the need of the user and the writing style of the authors. The searched information should be relevant to the user's query. However, there is often too much information related to a user query, for a user to browse.

Because information gathering plays a very important role, many researchers are delving into this area. A typical approach is to design a search engine. In the current market, search engines mainly fall into three types, keyword-based search engines, meta-search engines, and FAQ-based search engines. Most of current search engines are keyword-based, such as Yahoo and MSN. These engines accept a keyword-based query from a user and search in one or more index databases. They usually have huge databases of web sites that can be searched by inputting some text. Search engines index their information by sending out spiders or robots, which follow links from web sites and index all pages they come across. Each search engine has its own formula for indexing pages. Some index the whole site, while others index only the main page. Despite its simplicity, these engines typically return many thousands, even millions of sites in response to a simple keyword query, which often makes it impossible for a user to find the required information. For example, when we searched for "how to write a grant proposal", Google returned 366,000 sites, Yahoo returned 581,000 pages, and AltaVista returned 64,165 pages. The overloading is certainly a key problem for these search engines. Also, if you look at the first 50 pages from each search engine, the ranking is quite different due to the different ranking formulae. What we observe is that different search engines are good at different queries.

Based on the above analysis, the problem for our research can be formulated as follows. For a set of data sources from the Web, we are interested in *reducing irrelevant information by conditional association semantics*.

3. SIMILARITY MEASURES BY ASSOCIATION SEMANTICS

Let D be the set of terms in a given document, and Q be the set of terms in a query. There are two prevailing methods. One is based on terms of similarity, and another is based on latent semantics. Terms are considered independently in these models. In fact, all terms in D (or Q) have association semantics. In general, for any S the subset of D (or Q), and x in S, there is a semantics set of x, given S. This association semantics of terms should be considered in semantic indexing. We now present an approach for measuring similarity between two documents by latent semantics.

For a term t of D, the association semantics of t is a set of all possible semantics of t, denoted by $AS(t \mid D)$. That is,

$$AS(t \mid D) = \{\ s \mid s \ \text{ is a possible semantics of } t \text{ given } D\}$$

We define the distance between terms t_1 and t_2 of D below based on association semantics.

$$m_{AS}(t_1,t_2)=\frac{|AS(t_1\mid D)\bigcap AS(t_2\mid D)|}{|AS(t_1\mid D)\bigcup AS(t_2\mid D)|}$$

Example 1. Let t_1, t_2 and t_3 be three terms, and $AS(t_1 \mid D) = \{a_1, a_2, b_2, c_1\}$, $AS(t_2 \mid D) = \{a_2, b_1, b_2, c_1\}$, and $AS(t_3 \mid D) = \{a_1, a_2, b_2, c_1, c_2\}$. Then

$$m_{AS}(t_1,t_2)=\frac{|AS(t_1\mid D)\bigcap AS(t_2\mid D)|}{|AS(t_1\mid D)\bigcup AS(t_2\mid D)|}=\frac{3}{5}=0.6$$

$$m_{AS}(t_1,t_3)=\frac{|AS(t_1\mid D)\bigcap AS(t_3\mid D)|}{|AS(t_1\mid D)\bigcup AS(t_3\mid D)|}=\frac{4}{5}=0.8$$

$$m_{AS}(t_2,t_3)=\frac{|AS(t_2\mid D)\bigcap AS(t_3\mid D)|}{|AS(t_2\mid D)\bigcup AS(t_3\mid D)|}=\frac{3}{6}=0.5$$

The distance between a document D and a query Q can be then defined as follows, where $D = \{d_1, d_2, \ldots, d_n\}$ and $Q = \{q_1, q_2, \ldots, q_k\}$.

(1) The simplest similarity measurement is

$$M_{AS}(D,Q)=\frac{|(AS(d_1\mid D)\bigcup\cdots\bigcup AS(d_n\mid D))\bigcap(AS(q_1\mid Q)\bigcup\cdots\bigcup AS(q_k\mid Q))|}{|AS(d_1\mid D)\bigcup\cdots\bigcup AS(d_n\mid D)\bigcup AS(q_1\mid Q)\bigcup\cdots\bigcup AS(q_k\mid Q)|}$$

(2) For a rigorous similarity measurement, and without losing generality, we assume n ³ k. We construct the following distance table between terms.

In Table 1, $a_{ij} = m_{AS}(d_i, q_j)$ when $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$; $a_{ij} = 0$, when $i =1, 2, \ldots, n$ and $j = k+1, \ldots, n$.

We take the greatest value in the above as the distance between D and Q. That is,

$$M_{AS}(D,Q)=Max\{m_i\}_{i=1}^{N}$$

(3) The Boolean OR-Query, can be described in a standard format as a Boolean expression. The common Boolean expression is

$$Q = (q_1 \wedge \ldots \wedge q_i) \vee (q_{i+1} \wedge \ldots \wedge q_j) \vee \ldots \vee (q_{k+1} \wedge \ldots \wedge q_n)$$

Assume that $Q_1 = \{q_1, \ldots, q_i\}$, $Q_2 = \{q_{i+1}, \ldots, q_j\}$, …, and $Q_m = \{q_{k+1}, \ldots, q_n\}$. Then the query can be expressed as

$$Q = Q_1 \vee Q_2 \vee \ldots \vee Q_m$$

The similarity measurement between D and Q is defined as

$$M_{AS}(D, Q) = \text{Max}\{M_{AS}(D, Q_1), M_{AS}(D, Q_2), ¼,M_{AS}(D, Q_m)\}$$

***Table 1**: Mutual distances among terms given* D *and* Q

	q_1	q_2	…	q_k	$\varnothing$	…	$\varnothing$
d_1	a_{11}	a_{12}	…	a_{1k}	$a_{1(k+1)}$	…	a_{1n}
d_2	a_{21}	a_{22}	…	a_{2k}	$a_{2(k+1)}$	…	a_{2n}
…	…	…	…	…	…	…	…
d_n	a_{n1}	a_{n2}	…	a_{nk}	$a_{n(k+1)}$	…	a_{nn}

4. PROCEDURES FOR SIMILARITY CALCULATION

Because the similarity using latent semantics is similar to that of association semantics, we only present algorithms to compute the similarity of association semantics. Let D be a given document and Q be a query. We have,

```
Procedure 1.  SimpleSimMeasure
begin
Input: $D$: a document, $Q$: a query;
Output: $M_{AS}^{sim}(D,Q)$: the similarity;
(1) for $d \in D$ do
        begin
            generate $AS(d \mid D)$;
            let $AS_D \leftarrow AS_D \cup AS(d \mid D)$;
        end
        for $q \in Q$ do
        begin
            generate $AS(q \mid Q)$;
            let $AS_Q \leftarrow AS_Q \cup AS(q \mid Q)$;
        end
(2) let $M_{AS}^{sim}(D,Q) \leftarrow |AS_D \cap AS_Q| / |AS_D \cup AS_Q|$;
(3) output the similarity between $D$ and $Q$ is $M_{AS}^{sim}(D,Q)$;
endall.
```

The procedure *SimpleSimMeasure* estimates the similarity between two documents, D and Q, by using latent semantics.

An algorithm for calculation of the rigorous similarity of association semantics is given below, where, for simplicity, $D = \{d_1, d_2, \ldots, d_n\}$, $Q = \{q_1, q_2, \ldots, q_k\}$, and $n = k$.

```
Procedure 2. RigSimMeasure
begin
Input: $D$: a document, $Q$: a query;
Output: $M_{AS}^{rig}(D,Q)$: the similarity;
(1) input the weight set $\{w_1, w_2, \ldots, w_n\}$;
        for $d \in D$ do
            generate $AS(d_1 \mid D), AS(d_2 \mid D), \ldots, AS(d_n \mid D)$;
          for $q \in Q$ do
              generate $AS(q_1 \mid Q), AS(q_2 \mid Q), \ldots, AS(q_n \mid Q)$;
(2)  for $d \in D$ do
          for $q \in Q$ do
              let $a_{ij} \leftarrow m_{AS}(d_i, q_j)$;
(3) let $I \leftarrow$ the set of all possible reorders of $(1, 2, \ldots, n)$;
        let $M_{AS}^{rig}(D,Q) \leftarrow 0$;
        for $i = 1$ to $n$ do
          for any $(l_1, l_2, \ldots, l_n) \in I$ do
          begin
            let $tem \leftarrow w_1 * a_{il1} + w_2 * a_{il2} + \ldots + w_n * a_{iln}$;
            if $tem > M_{AS}^{rig}(D,Q)$ then
                let $M_{AS}^{rig}(D,Q) \leftarrow tem$;
          end
    (4) output the similarity between $D$ and $Q$ is $M_{AS}^{rig}(D,Q)$;
    endall.
```

The procedure *RigSimMeasure* estimates the similarity between two documents, D and Q, by using association semantics.

5. COMPARISON AND SUMMARY

For convenience, our comparison is only focused on the simplest formulae of conventional similarity measurement $M_{pre}(D, Q)$, the similarity measurement by latent semantics $M_{LS}(D, Q)$, and the similarity measurement by association semantics $M_{AS}(D, Q)$. We have

$$M_{pre}(D,Q) = \frac{|D \bigcap Q|}{|D \bigcup Q|}$$

$$M_{LS}(D,Q) = \frac{|(LS(d_1) \bigcup \cdots \bigcup LS(d_n)) \bigcap (LS(q_1) \bigcup \cdots \bigcup LS(q_k))|}{|LS(d_1) \bigcup \cdots \bigcup LS(d_n) \bigcup LS(q_1) \bigcup \cdots \bigcup LS(q_k)|}$$

$$M_{AS}(D,Q) = \frac{|(AS(d_1 | D) \bigcup \cdots \bigcup AS(d_n | D)) \bigcap (AS(q_1 | Q) \bigcup \cdots \bigcup AS(q_k | Q))|}{|AS(d_1 | D) \bigcup \cdots \bigcup AS(d_n | D) \bigcup AS(q_1 | Q) \bigcup \cdots \bigcup AS(q_k | Q)|}$$

Suppose $D = \{d_1, d_2, d_3\}$ = {*discovery*, *data set*, *knowledge*}, and $Q = \{q_1, q_2\}$ = {*mine*, *rule*}. Certainly, we have

$$M_{pre}(D,Q) = \frac{|D \bigcap Q|}{|D \bigcup Q|} = 0$$

In order to apply $M_{LS}(D, Q)$ and $M_{AS}(D, Q)$, assume $LS(d_1)$ = {*discovery*}, $LS(d_2)$ = {*data set*, *database*}, $LS(d_3)$ = {*knowledge*}, $LS(q_1)$ = {*mine*, *belonging to me*}, $LS(q_2)$ = {*rule*} and $AS(d_1|D)$ = {*discovery*}, $AS(d_2|D)$ = {*dataset*, *database*, *document set*}, $AS(d_3|D)$ = {*knowledge*, *rule*, *law*, *data*}, $AS(q_1|Q)$ = {*mine*, *discovery*, *extraction*, *learning*}, $AS(q_2|D)$ = {*rule*, *knowledge*, *law*}. Then, we have

$$M_{LS}(D,Q) = \frac{|(LS(d_1) \bigcup \cdots \bigcup LS(d_n)) \bigcap (LS(q_1) \bigcup \cdots \bigcup LS(q_k))|}{|LS(d_1) \bigcup \cdots \bigcup LS(d_n) \bigcup LS(q_1) \bigcup \cdots \bigcup LS(q_k)|} = 0$$

$$M_{AS}(D,Q) = \frac{|(AS(d_1 | D) \bigcup \cdots \bigcup AS(d_n | D)) \bigcap (AS(q_1 | Q) \bigcup \cdots \bigcup AS(q_k | Q))|}{|AS(d_1 | D) \bigcup \cdots \bigcup AS(d_n | D) \bigcup AS(q_1 | Q) \bigcup \cdots \bigcup AS(q_k | Q)|} = \frac{4}{10} = 0.4.$$

As we have seen, with the explosive growth of information on the WWW, there is a great need for efficient information searching relevant to user queries. By using search engines, such as Yahoo, MSN, and Google, many thousands, even millions of results are usually returned in response to a user query. It would be difficult for a user to browse so much searched information. In particular, it is an important challenge to identify which pieces of the information are really useful to the user. In this paper, we designed a new strategy for information indexing by conditional association semantics. The proposed approach can efficiently reduce irrelevant information searched.

6. REFERENCES

[1] C. H. Chang and C. C. Hsu. Enabling concept-based relevance feedback for information retrieval on the WWW. *IEEE Transactions on Knowledge and Data Engineering*, 1999, 11(4): 595-609.

[2] N. R. Jennings, K. Sycara, and M. Wooldridge. A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems*, 1998, 1(1): 7-38.

[3] V. Lesser, B. Horling, F. Klassner, A. Raja, T. Wagner, and S. Zhang. A next generation information gathering agent. *Proceedings of the 4th International Conference on Information Systems, Analysis, and Synthesis*, Orlando, FL, July 1998.

[4] V. Lesser, B. Horling, F. Klassner, A. Raja, T. Wagner, and S. Zhang. BIG: An agent for resource-bounded information gathering and decision making. *Artificial Intelligence Journal, Special Issue on Internet Information Agents*, Vol. 118, 1-2(2000): 197-244.

[5] S. Li and P. Danzig. Boolean similarity measures for resource discovery. *IEEE Trans. Knowledge and Data Eng.*, vol. 9, 6(1997): 863-876.

[6] X. Wang, Shichao Zhang, P. K. Khosla, H. Kiliccote and Chengqi Zhang. Anytime algorithm for agent-mediated merchant information gathering. *Proceedings of the Fourth International Conference on Autonomous Agents*, Catalonia, Spain, ACM Press, June, 2000: 333-340.