The Strategy of Matching User Queries with Web Pages Based on Formal Concept Analysis

YaJun Du\textsuperscript{1}, HaiMing Li\textsuperscript{2}, and Zhenwen Liao\textsuperscript{3}

\textsuperscript{1} School of Mathematics and Computer Engineering, Xihua University, Chengdu 610039, China
\textsuperscript{2} School of Information Science and Technology, Zhe Jiang Forestry University, Lin’an 311300, China
\textsuperscript{3} Chengdu Center, China Geological Survey, Chengdu 610081, China

Abstract. A searching result can be returned for submitting user query to Search Engine (SE). However, large numbers of Web pages in the searching result are not interested by users. To overcome this problem, from automated reasoning perspective, we put forward a matching model how to match user query with Web pages. According as the basic theory of formal concept lattice, we defined OR-RULE and AND-RULE, put forward AND-MATCH and OR-MATCH. Two reasoning methods based on AND-MATCH and OR-MATCH, and their algorithms of the matching model, are proposed. The experiment demonstrates that these method is efficient.

1 Introduction

A searching result can be returned for submitting user query to Search Engine (SE). However, large numbers of Web pages in the searching result are not interested by users. To solve this problem, matching user query with Web pages has become a very important study. Salton et al. (1975) presented the vector space model. Raghavan (1986); Salton et al. (1993) represents the document and user query into n dimension vector \(((t_1, w_1), \ldots, (t_n, w_n))\) separately, where \(t_i\) is a key word (term), \(w_i\) is the frequency of \(t_i\), the matching distance between user query and Web pages can be computed. Bookstein (1980) reported on the boolean model of information retrieval. Gerard et al. (1983); Salton (1991); Crouch et al. (1989); Bordogna & Pasi (1993) regarded the boolean model as the matching model. If keyword \(t_i\) in user query is contained in Web page, then \(w_i\) is assigned into True, else \(w_i\) is assigned into False. Gerard et al. (1983) showed the retrieval system of the extended boolean model. Some studies (Kwok, 1990; Robertson et al., 1996) demonstrated that probability model is better than the boolean model and the vector space model. Researchers have recently begun to develop the semantic matching model. Deerwester et al. (1990) proposed the latency Semantic index (LSI) model. Losada & Barreiro (2001); Losada & Luna (2003) discussed the logic model of proposition between user query and Web pages. Rajapakse & Denham (2003, 2005) put forward the bidirectional associative memory, they combined formal concept analysis (FCA) with bidirectional associative memory (BAM) to research on matching user query with Web pages.
2 Our Related Works

In our Personalized Intelligent Search Engine (PISE)\cite{Du et al., 2005a, Li et al., 2004}, spiders are developed to automatically collect related Web pages whenever users submit their query $Q$ (a set of some key words $Q = \{k_1, \cdots, k_n\}$). PISE filtrates some of the related Web pages and returns the results (some Web pages, denote $R_Q$) in which every Web pages is useful to user.

Definition 1. Let $R_Q, \overline{R_Q}$ are LOW-ADJACENCE and UPPER-ADJACENCE set of $R_Q$ separately, they are defined as follows:

$$R_Q = \{v_i | v_i \in V, \frac{|\pi(v_i) \cap Q|}{|Q|} = 1\}, \overline{R_Q} = \{v_i | v_i \in V, \frac{|\pi(v_i) \cap Q|}{|Q|} \leq 1\}.$$  

where for all $v_i \in V$, $\pi(v_i)$ is key word set of $v_i$.

2.1 Construct Concept Lattice of LOW-ADJACENCE Set of $R_Q$

Definition 2. A formal context of $R_Q$ is a triple $(R_Q, D, P)$, where $R_Q = \{v_1, \cdots, v_n\}$, each $v_i (i \leq n)$ is called an object, and $D = \bigcap_{v_i \in V} \pi(v_i)$ is the common key word set of all Web pages in $V$, $D = \{d_1, \cdots, d_m\}$, each $d_j (j \leq m)$ is called an attribute. Finally, let $P = \{p_j : j \leq m\}$, $p_j : R_Q \rightarrow D_j$, $p_j$ is the mapping function between $R_Q$ and $D_j$ ($D_j$ is the value domain of attribute $d_j$).

In real application, $D_j$ is relational to TF-IDF(Term Frequency, Inverse Document Frequency), $(R_Q, D, P)$ is a continuous-valued formal context. Discretization of $D_j$ is needed before constructing its the concept lattice \cite{Du et al., 2005a}.

Example 1. There are seven Web pages labeled 1, 2, 3, 4, 5, 6, 7, separately. Let $Q = \{A, B\}$, $V = \{1, 2, 3, 4, 5, 6, 7\}$, $D = \{A, B, C, E\}$, $R_Q = \{1, 2, 3, 4\}$, $\overline{R_Q} = \{1, 2, 3, 4, 5, 6, 7\}$. TF-IDF score of every key words in every Web pages is calculated and listed in Table I (left). We discretize $R_Q$ into many-valued formal context (Table I (right)) by 4 intervals of every attribute \cite{Du et al., 2005a}.

\begin{table}
\centering
\begin{tabular}{lcccc}
A & B & C & E \\
1 & 0.50 & 0.20 & 0.05 & 0.25 \\
2 & 0.08 & 0.20 & 0.30 & 0.42 \\
3 & 0.46 & 0.08 & 0.11 & 0.35 \\
4 & 0.15 & 0.55 & 0.10 & 0.20 \\
5 & 0.30 & 0.00 & 0.20 & 0.50 \\
6 & 0.00 & 0.45 & 0.35 & 0.20 \\
7 & 0.00 & 0.00 & 0.40 & 0.60 \\
\end{tabular}
\caption{$Q = \{A, B\}$ and its Formal Context of $R_Q$}
\end{table}