The Optimization in News Search Engine Using Formal Concept Analysis

Yan Liu, QingXian Wang, Lei Guo, Qing Yao, Nan Lv and Qiang Wang
Information Engineering Institute, Information Engineering University
liu_yan_hello@yahoo.com.cn

Abstract

Being faced with the huge amount and rapid renewal of the news data, how to help users to obtain the news content they need rapidly, and how to use the resource of news information effectively, both come to the present problems. Formal Concept Analysis(FCA) arose as a mathematical theory for the formalization of the concept of 'concept' and has more advantages in concept description and relation measurement than binary-relationship such as database. The technique of using FCA to optimize the personal news search engine is proposed in this paper, including users' query background construction using FCA , query keywords optimization based on users' background and the new layout strategy of search result based on "Concept Tree".

1. Introduction

News browsing and searching is one of the most important Internet activity. The huge amount of news available online reflects the users' need for a plurality of information and opinions. News Search engines are then a direct link to fresh and unfiltered stream of information. There are many commercial news search engines, for example Google News, Yahoo news, Microsoft NewsBot, etc.

Despite this great variety of commercial solutions, we found just few academic research on this subject. Being faced with the huge amount and rapid renewal of the news data, how to help users to obtain the news content they need rapidly, and how to use the resource of news information effectively, both come to the present problems.

The rest of the paper is organized as follows. In Section 2, related works is discussed and current problems of the present news search engines are analyzed. In Section 3 the theory of Formal Concept Analysis is introduced. The technique of using FCA to optimize the personal news search engine is discussed in detail in Section 4. Finally, we conclude our work.

2. Related Works

Nowadays, as the primary means of information retrieval, commercial news search engines have basically solved the problem of collecting huge arsenals of news and fetching useful information to news readers. But they couldn't satisfy readers in some fields, such as the coverage of resources, the accuracy of result and the clear of visualization.

There are some works related to improving news search engine. NewsInEssence [1] is a system for finding and summarizing clusters of related news articles. Chung [2] proposes a topic mining framework for news data stream. Henzinger [5] finds news articles on the web that are relevant to TV news currently being broadcast. Reis [7] proposes a tool to automatically extracting news from Web sites. NewsJunkie [3] is a system that personalizes news for users by identifying the novelty of stories in the context of stories users have already reviewed.

Although these efforts have been made, there are still many deficiencies in the news search services need to be improved. The News search engine technology remains largely a black art.

3. Theoretical Basis

Formal Concept Analysis [4] arose as a mathematical theory for the formalization of the concept of 'concept' in the early 80ies and is nowadays considered as an AI theory. It has since then grown to a technique for data analysis, information retrieval, and knowledge representation with over 200 applications.

In this section, we briefly recall the basic notions of Formal Concept Analysis.

Definition 1

A formal context is a triple $\mathrm{K} := (G, M, I)$, where G and M are sets , and $I \subseteq G \times M$ is a binary relation. The elements in G are called objects, and the elements in M are called attributes. Using gIm or $(g, m) \in I$ to express the relationship between object g and attribute m, which is called "object g has attributes m" .

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE

Definition 2

For objects subset A which belongs to set G,$A \subseteq G$, there is set $A' := \{m \in M | (g, m) \in I, \forall g \in A\}$ which represents the attributes' set of all the objects in A.

For attributes subset B, $B \subseteq M$, there is set $B' := \{g \in G | (g, m) \in I, \forall m \in B\}$ which represents the objects' set related to all the attributes in B.

Definition 3

A formal concept which belongs to formal context (G, M, I) is a set of pair (A, B), $A \subseteq G, B \subseteq M$, satisfying $A' = B$ and $B' = A$. A, B are the extention and intention of the formal concept.

Definition 4

Let the two formal concepts $(A1, B1), (A2, B2)$, if $A1 \subseteq A2$(equals to $B2 \subseteq B1$), $(A1, B1)$ is called the sub-concept of $(A2, B2)$, and $(A2, B2)$ is called the super-concept of $(A1, B1)$, namely $(A1, B1) \leq (A2, B2)$. Relation " $\leq$ " is the sequence of formal concept, which reflects the hierarchy in the concepts. Thus, the ordered set of all the formal concepts of (G, M, I) is expressed as $\beta(G, M, I)$, which is called the concept lattice of formal context (G, M, I).

The construction of concept lattice is a process of clustering the concepts. The result lattice of the concept is the only one from the same data, which is not influenced by the sequence of data or attributes.

4. Optimization in News Searching Using FCA

In this section, we apply the FCA to news information retrieval, trying to make the most use of the concept lattice to express the relationships of concepts, and make some breakthroughs in query keywords optimization and the results visualization.

4.1. Data expression

The locations, characters, organizations etc. are the important metadata for event description. We should pay more attention to these metadata and catch the key elements to exhibit the basic elements of events. The elements of news could be simplified as the following four basic sets:

(1) Character set: $P = \{p_1, p_2, p_i, \cdots, p_m\}$

(2) Organizations set: $O = \{o_1, o_2, o_i, \cdots, o_l\}$

(3) Locations set: $L = \{l_1, l_2, l_i, \cdots, l_t\}$

(4) Events set: $E = \{e_1, e_2, e_i, \cdots, e_n\}$.

Let $W = P \cup O \cup L$, the initial attributes set of e_i is W_i, $W_i \subseteq W$, each event e_i is represented as w-dimensional vector $e_i(p_{i1}, \cdots, p_{im}, o_{i1}, \cdots, o_{il}, l_{i1}, \cdots l_{it}), w = |W_i|$, Thus, we obtain the vector space model of Events :V_E.

$$V_E = \begin{pmatrix} e_1, & \cdots & e_i, & \cdots & e_n \end{pmatrix}^T \tag{1}$$

Here we take the event of news as the concepts of formal concept analysis and the metadata of news (character names, organization names, location names, etc.) as attributes.

4.2. Concept Acquisition

The process of obtaining concepts from news information is shown in figure 1.

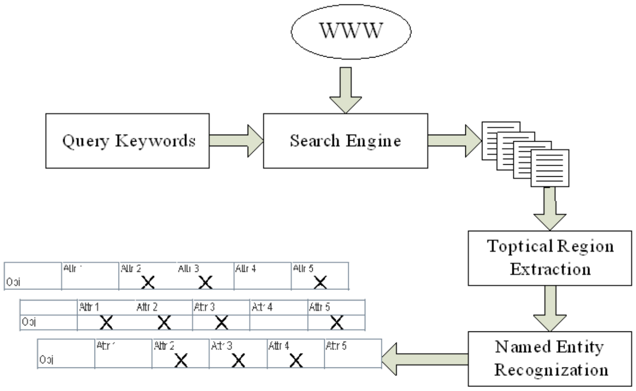


Figure 1. The Process of Concept Acquisition

step 1 Collect related pages set according to event e_i.

The page set is expressed by $Page = \{u_1, u_2, u_i, \cdots, u_N\}$, where N is constant, representing the top N results of the searching records. As the normal search engines tend to have low accuracy and recall, we expand the scope of results using meta-search engine, combining three-layer structure mining algorithms to improve the accuracy of search result [8] .

step 2 Obtain the topical blocks from each page u_i.

According to the method of [6], We adopt a vision-based page segmentation algorithm to partition a Web page into semantic blocks and use spatial features and content features to represent each block. Shannon's information entropy is adopted to represent the information strength of each feature and an entropy-based improved Naïve Bayes classifier is used to identify primary area within Web pages. Thus, the topical block $topicalblock_i$ of each page is extracted.

step 3 Extract named entities from $topicalblock_i$ to construct the attributes set of concepts.

The metadata is extracted from the news using the technique of Named Entity Recognition. As a result, we obtain the sets of character names (P_i), organization names (O_i) and location names (L_i), which are associated with the event e_i. The set $W_i = P_i \cup O_i \cup L_i$ contains all kinds of metadata of top N pages returned.

Currently, Named Entity Recognition technology has achieved perfect success, which provides technical assur-

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE

ance for getting news' metadata. We choose the research results of Stanford Natural Language Processing Center in our work.

4.3. Users' Background Construction

Users' background concerns users' query history. The construction of users' background is related to two aspects: the one is to obtain the relationships between attributes in each event. Another is to analyze relationships between several events.

(1)News Features Extraction

Considering that not all the metadata in named entity sets P_i, O_i, L_i are closely relevant to the subject. It is necessary to take into account the filtration problem to compress the features set of news. If we take each metadata in sets P_i, O_i, L_i as the features of event e_i, the filter issues of unrelated metadata would be converted into a traditional problem of feature selection.

Therefore, we use "feature frequency factor" to measure the weight of features which represents the frequency of each metadata occurs in the result pages. And feature selection threshold $\varphi_p, \varphi_o, \varphi_l$ decide whether the word W_i is the feature of event e_i or not, the subset W_i', $W_i' \subseteq W_i$ is the last attributes set of event e_i.

The ultimate metadata set is the named entities that are the most frequently occur in the searching result. Taking this set as attributes set, we get the relationship between event and attributes, which is used as the basic formal context for the next analyzing.

(2)Event-Attributes Set Acquirement

For each event e_i that has been queried in the history, extract the news metadata to get the event's feature set W_i'. Assuming the number of query is M in the user's query history, events' attribute set associated with all the related events are defined as a set : $W = W_1' \cup W_2' \cup W_3' \cup \cdots W_M'$.

(3) User Background Construction

If the events are taken as objects and the news' metadata associated with the event are taken as attributes, the users' formal background could be constructed, namely $K := (E, W, I)$, in which $(e_i, w_j) \in I$ means the feature w_j is a metadata that is related with the event e_i, $0 \leq i \leq M, 0 \leq j \leq |W|$.

4.4. The Optimization in Query Keywords

Traditional keywords-based search engines tend to return inadequate results that are accord with users' requirement. One reason is that the keywords users input is too few that search engines may return many irrelevant documents, and lead to low accurate rate. The second one, due to lacking enough knowledge of the query content, user could not give accurate keywords, which leads to low recall rate. Therefore, It is necessary to improve the query quality and optimize the organization of the searching result.

1. The Framework of Implement

If the search engine could provide some suggestions associated with the keywords, the process of query would be more effective. Therefore, the optimization of keywords is important for improving the quality of search results. It could be considered from two aspects:

(1)The expansion of keywords, which is several '$\cup$' operations on the query keywords;

(2)The limitation of keywords, which is several '$\cap$' operations on the query keywords;

The expansion and limitation depends on the understanding of keywords' concepts. Therefore, it is very importance to calculate the similarity between concepts and select the similar concepts as additional keywords to expand or limit the query. Moreover, the optimization of query keyword is helpful for improving the recall and accurate rate in search engines. The framework of implement is shown in Figure 2.

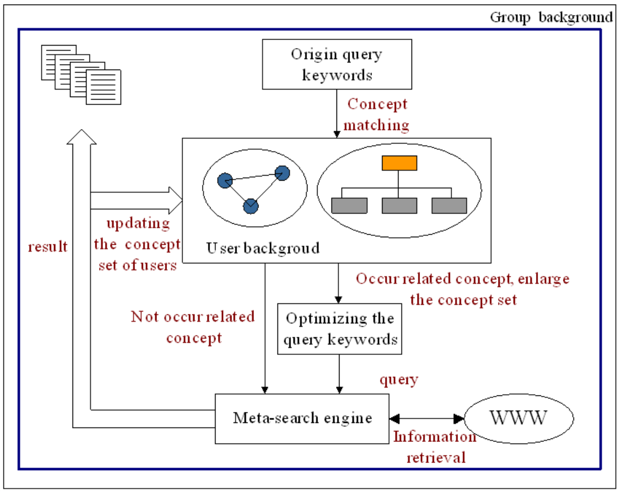

Group background
Origin query keywords
Concept matching
updating the concept set of users
result
User backgroud
Occur related concept, enlarge the concept set
Not occur related concept
Optimizing the query keywords
query
Meta-search engine
WWW
Information retrieval

Figure 2. Information Retrieval Framework

Here users' background guides the selection and expansion in the process of news retrieval. In order to correctly understand users' interest, the expansion of query keywords should consider the following factors:

(1) the construction of users' background;

(2) the similarity measurement between concepts;

For (1), from the viewpoint of formal context, users' formal context describes the relationships among concepts and the relationships between the concepts and their attributes. These relations and their expansions are not to list the keywords simply, but to describe their inherent relationships reasonably by network structure.

For (2), the calculations of similarity between concepts are the core of query keywords expansion and searching results sorting. Based on the theory of concept lattice, we utilize the attributes' correlations between concepts to measure similarity between the keywords in user's background.

2. Similarity Calculation Based on Concept Lattice

Formal concept analysis can express the concepts, the attributes and the relationships reasonably using formal lattice. And the exclusive correlative concept lattice constructed could express the structure of concept clearly and help to find out the possible concepts and classification relationship between concepts.

Therefore, the similarity between concepts can be calculated using concepts lattice. Let C_1 and C_2 are two concepts, the formal context is $K = (G, M, I)$. In formal context, C_1 and C_2 are taken as two elements in the objects sub-set A, $A = \{C_1, C_2\} \subseteq G$. If C_1, C_2 have multi-valued attributes, the formal context should be converted to single-valued attributes. The ultimate single-valued formal context is $K' = (A, M', I')$.

Based on the formal concept analysis, we get the concepts lattice containing C_1 and C_2, $Sim(C_1, C_2)$ represents the similarity between these two concepts, which can be calculated as:

$$Sim(C_1, C_2) = \frac{|A'|}{N} \tag{2}$$

in which $A' = \{m \in M | (g, m) \in I, \forall g \in A\}$ is the shared attributes set in C_1 and C_2. N is the height of the concept lattice, which is the maximum number of attributes of single object in formal context.

Each node in concept lattice represents a concept. Formally, the description of a concept contains two aspects which are extension and intension. The extension is the set of all objects which belongs to this concept, and the intension is the sharing attributes set of all the objects which belong to the concepts.

Usually, the number of the objects in concept extension may be more than one, so that it is easy to utilize the concept extension for expanding similar keywords. When one object which belongs to the extension is selected, it can be concluded that the other objects which belong to the same extension should be included in the result because of the same concept that these objects contain and the same attribute that these objects share. In this way, using the idea of concept extension could improve the scope of searching greatly, and there are inherent relationships among the search results. The process is described in Figure 3.

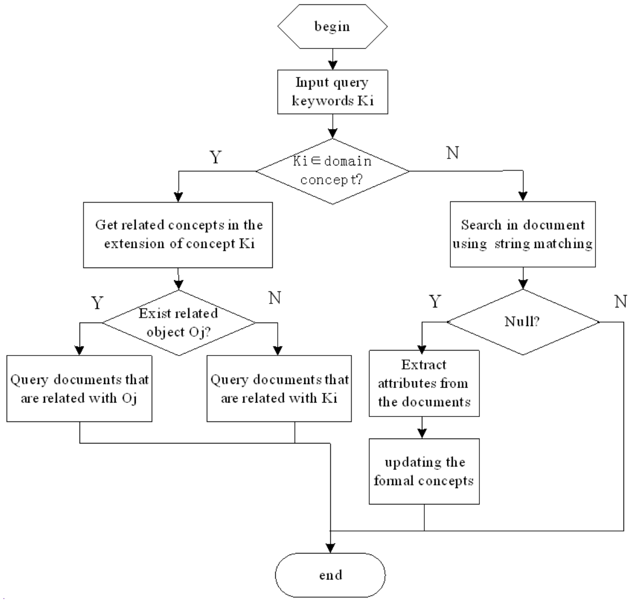

begin
Input query keywords Ki
Ki∈domain concept?
Y
N
Get related concepts in the extension of concept Ki
Search in document using string matching
Exist related object Oj?
Y
N
Y
N
Null?
Query documents that are related with Oj
Query documents that are related with Ki
Extract attributes from the documents
updating the formal concepts
end

Figure 3. Process of Concept Similarity Calculation

4.5. The Optimization in Visualization

Usually users hope to explore the query results that are categorized and ordered according to the relevance or authority and catch the key points of the news quickly. News is a special data source. It is well written and emphasis on format. Characters, organizations and locations are the three important metadata that can summarize the topical meaning of the news. Thus, if we can make good use of these metadata, we could catch the topic quickly.

FCA describes the relationship among objects and their attributes. If we take the news pages as objects and their metadata as their attributes, the news pages that are related with the query keywords could be organized by the formal concept. And the metadata could reflect the topic of a page from some perspective. Therefore, it is a better idea to solute the problem of organizing a large number of results in more hierarchical layout and reducing the cost of calculation about semantics.

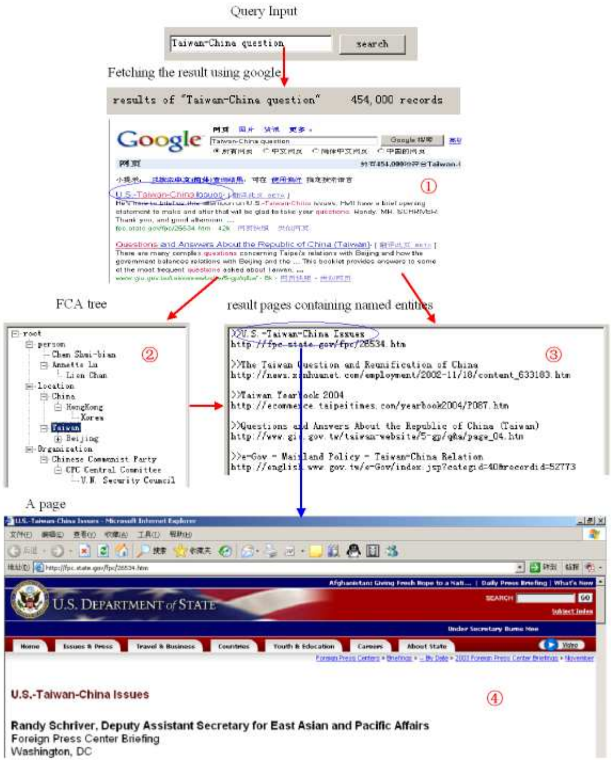
Figure 4 shows the main steps of our news search engine using FCA to help organizing the searching result.

Step 1 User inputs a set of keywords $W = \{w_1, w_2, \cdots, w_n\}$. The set of recalled pages is $P = \{p_1, p_2, \cdots, p_N\}$. Considering that user tend to pay more attention to higher rank pages, we select only the Top N pages as the source for analyzing.

Step 2 Extract metadata from each page p_i. As a result, we get the named entities sets, $U_i = \{u_{i1}, u_{i2}, \cdots, u_{iN_i}\}$, N_i is the number of named entities in page p_i.

All the named entities of records set P is the set U, namely $U = U_1 \cup U_2 \cup U_i \cup \cdots U_N$.

Step 3 Generating concept lattice of HTML pages, in which pages are objects and metadata are attributes. The formal context is represented as $K := (P, U, I)$, where



U.S.-Taiwan-China Issues
He's here to brief us this afternoon on U.S.-Taiwan-China issues. He'll have a brief opening statement to make and after that will be glad to take your questions. Randy. MR. SCHRIVER: Thank you, and good afternoon. ...
fpc.state.gov/fpc/26534.htm

Questions and Answers About the Republic of China (Taiwan)
There are many complex questions concerning Taipei's relations with Beijing and how the government balances relations with Beijing and the ... This booklet provides answers to some of the most frequent questions asked about Taiwan. ...

```
root
  person
    Chen Shui-bian
    Annette Lu
      Lien Chan
  location
    China
      HongKong
        Korea
      Taiwan
        Beijing
  Organization
    Chinese Communist Party
      CPC Central Committee
        U.N. Security Council
```

```
>>U.S.-Taiwan-China Issues
http://fpc.state.gov/fpc/26534.htm

>>The Taiwan Question and Reunification of China
http://news.xinhuanet.com/employment/2002-11/18/content_633183.htm

>>Taiwan Yearbook 2004
http://ecommerce.taipeitimes.com/yearbook2004/P087.htm

>>Questions and Answers About the Republic of China (Taiwan)
http://www.gio.gov.tw/taiwan-website/5-gp/q&a/page_04.htm

>>e-Gov - Mainland Policy - Taiwan-China Relation
http://english.www.gov.tw/e-Gov/index.jsp?categid=40&recordid=52773
```

U.S.-Taiwan-China Issues - Microsoft Internet Explorer
http://fpc.state.gov/fpc/26534.htm
Afghanistan: Giving Fresh Hope to a Nati... | Daily Press Briefing | What's New
Under Secretary Burns Hee
Foreign Press Centers » Briefings » ... By Date » 2003 Foreign Press Center Briefings » November

U.S.-Taiwan-China Issues

Randy Schriver, Deputy Assistant Secretary for East Asian and Pacific Affairs
Foreign Press Center Briefing
Washington, DC

Figure 4. Visualization of Searching Results

$(p_i, u_j) \in I$ means named entity u_j occurs in page p_i, $0 \le i \le N, 0 \le j \le |U|$.

Layout conversion is a effective method for vast data visualization and the tree has been proved that it can be completed in polynomial computation. Thus, for formal context $K := (P, U, I)$, the concepts' lattice could be transferred into tree-structure for search result exhibition, which is called "Concepts Tree", where root node means the supremum of the concept lattice and the relationships of nodes means the level of relations in the lattice. As shown in the figure 4, the higher the level of entities is, the wider coverage of the searching results is. Every node in the tree contains the abstract and the page URL.

Step 4 Navigating the URL to its Web site.

Base on the formal background, the concept lattices constructed by Web pages and named entities is based on the characteristics of concepts' lattices. On one hand, the structure of the concept hierarchy directly shows the relationships between concepts generally and specially. On the other hand, it shows the differences roles of named entities in the query results.

Concept lattice does not denote the real physical relations, such as the containing relationship between areas. Instead, it shows the relationships of concepts' hierarchy that is brought forward by the query background. This relationship exhibits the results from the perspective of semantic.

5. Conclusion

The analysis of current news search engines show that the services are not enough, the ability of individuation is still poor and the organizations of the results is yet not reasonable. This paper is based on the status of current news search engines. The current problem is analyzed, the theory of formal concept analysis is utilized in news search service, and the users' background is constructed, which forth more optimizes the query keywords and visualization of the searching results.

The application shows that it is helpful to improve the quality of present news search services. Of course, the extraction of the named entities and the generations of the concept lattice both bring the system an extra calculating cost. Some strategies are introduced to reduce the calculation costs such as limiting the number of pages returned and filtering the redundant data from the Web pages, etc. In the next work of research, there is still much that can be done to improve the performance of news search engine.

References

[1] S. Blair-Goldensohn and D. R. R. et al. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Proceedings of the first international conference on Human language technology research[C]*, pages 1–4, San Diego, 2001.

[2] S. Chung and D. McLeod. Dynamic topic mining from news stream data. In *Proceedings of ODBASE[C]*, pages 653–670, Catania, Italy, 2003.

[3] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th conference on World Wide Web[C]*, pages 482–490, New York, USA, 2004.

[4] B. Ganter and R. Wille. Formal concept analysis: Mathematical foundations[m]. *Springer Heidelberg*, 1999.

[5] M. Henzinger and B. C. et al. Query-free news search. In *Proceedings of the 12th international conference on World Wide Web[C]*, pages 1–10, Budapest, Hungary, 2003.

[6] Y. Liu, Q. Wang, and Q. Wang. A heuristic approach for topical information extraction from news pages. In *WISE2006, LNCS4255[C]*, pages 357–362, Wuhan,China, 2006.

[7] D. Reis and P. G. et al. Automatic web news extraction using tree edit distance. In *Proceedings of the 13th conference on World Wide Web[C]*, pages 502–511, New York, USA, 2004.

[8] Q. Wang, Y. Liu, and J. Luo. Exploiting link analysis with a three-layer web structure model. In *WISE2006, LNCS4255[C]*, pages 187–198, Wuhan,China, 2006.

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE