Association Thesaurus Construction for Interactive Query Expansion based on Association Rule Mining

Hahn-Ming Lee[1], Chi-Chun Huang[2] and Chun-Yen Chao[3]
[1] Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
hmlee@mail.ntust.edu.tw
[2] Department of Information Management
National Kaohsiung Marine University
cchuang@mail.nkmu.edu.tw
[3] Department of Electronic Engineering
National Taiwan University of Science and Technology
lowson@interepoch.com.tw

Abstract- *This paper presents an interactive query expansion method with association thesaurus, which is mined from the 'selected web pages' of users in the query logs. The 'selected web pages' of users are transferred into 'sets of query terms' and then used for term correlation mining. Accordingly, various association thesauruses concerning different query terms are constructed from these term correlations. Consequently, the proposed method combines the original query term specified by a user with the corresponding thesaurus to offer the user more precise results. The query expansion mechanism is implemented within the Query Agent of a course recommendation system, Coursebot. Experimental results have shown that the performance, precision ratio and recall ratio, of the system is increased when the proposed method is applied.*

Keywords: association thesaurus construction, interactive query expansion, association rule mining, user feedback, Coursebot.

1. Introduction

With a rapidly growing number of Internet users, the World Wide Web (WWW) has increasingly become a primary medium for information dissemination and acquisition in recent years [22]. Nowadays, numerous websites provide people to surf the Internet with a huge amount of contents and hyperlinks. Owing to information overload, however, an appropriate tool to help users finding the desired information is necessary. Consequently, different approaches for effectively and efficiently discovering information and resources on the WWW have been investigated [20][18][19], such as a well-known technique—search engines.

By querying a search engine [20][18][19], a user probably can get what he or she wants. However, the average number of terms (words) specified by a user in a query is generally two to three [4][14]. This often leads to some problems. First, the search engine may respond a lot of irrelevant results. Furthermore, it is usually difficult for a user to describe his or her request precisely according to just a few query terms (words). To overcome these problems, various query expansion (or modification) techniques have been developed [2][8][7][11][16]. That is, the search engine returns more precise results to a user based on a refined word list, which is obtained by expanding (or modifying) the original query term or terms.

In general, two kinds of methods are commonly used for query expansion: (a) Automatic query expansion [11]; (b) Interactive query expansion [2][8][7][16]. In automatic query expansion (AQE), a user's query is expanded automatically according to some useful information, such as co-occurrence data, document classification, syntactic context, and so on [11]. In interactive query expansion (IQE), by contrast, a user has to select some additional search terms and then the original query is expanded based on these selected terms. Another variety of IQE is relevance feedback [12]. In this method, a user has to identify relevant documents in an initial retrieved document set. Then a new query is created based on these relevant documents.

Moreover, thesaurus has attracted great attention for query expansion (formulation) in recent years, such as synonym-based thesaurus [3][9] and similarity thesaurus [5][6][10][14][15]. To create a

synonym-based thesaurus, a set of synonym terms should be identified from a dictionary of words. However, term co-occurrence data are not considered within the synonym-based thesaurus. As for similarity thesaurus [10], term similarity (term-to-term relationship, such as term co-occurrence) is determined and then the thesaurus for a term i is created from terms with high similarity to term i.

To expand a query effectively, we believe that the query logs of users are useful. That is, the results of a query can be refined precisely when useful information is derived from the query logs containing *selected web pages* (or documents) of users. When a user enters a query term, he or she will be presented with search results (for example, web pages or documents). Accordingly, some of these web pages (or documents) are selected by the user. Consequently, the entire information about the entered query term, such as selected web pages, user id, and query term, is recorded in the query logs. For example, a query term 'A' can be expanded with term 'B' according to the high correlation between terms 'A' and 'B' in selected web pages of users. Such useful information can be viewed as the term co-occurrence data, which can be mined from the selected web pages of users. Thus, this paper presents an interactive query expansion method based on association thesaurus, which is created by mining association rules [1] from a set of query terms. The selected web pages (or documents) of users in the query logs are transferred into 'sets of query terms' and then used for term correlation mining. Accordingly, various association thesauruses concerning different query terms are constructed from these term correlations. Consequently, the proposed method combines the original query term specified by a user with the corresponding thesaurus to offer the user more precise results.

This approach takes advantages of both automatic and interactive query expansion methods because the query is expanded automatically and users are involved implicitly for query expansion. Moreover, the generated association thesauruses offer properties of both similarity thesauruses and synonym-based thesauruses, such as term co-occurrence and synonym-like terms determined implicitly by users.

The query expansion mechanism is implemented within the Query Agent of a course recommendation system, Coursebot [21]. Experimental results have shown that the performance, precision ratio and recall ratio, of the system is increased when the proposed method is applied.

The rest of this paper is structured as follows. Section 2 briefly reviews the concept of association rule mining. Section 3 gives a framework to describe the proposed approach used for constructing association thesaurus. Section 4

presents the Query Agent of a course recommendation system, Coursebot, within which the proposed query expansion method is implemented. Experimental results are reported in Section 5. Finally, Section 6 concludes.

2. Association rule mining

This section briefly reviews the concept of association rule mining adopted for constructing association thesaurus. Mining association rules from a large collection of data has gained great popularity in information retrieval since it was proposed in 1993 [1]. Similar to decision rules, association rules are used to describe the relationships between sets of items. Let I be a set of items (binary attributes). An association rule is generally stated as the following expression [1]:

$$X \rightarrow Y | c, \qquad (1)$$

where $X \subset I$, $Y \subset I$, $X \cap Y = \phi$, and c is a constant indicating the confidence of the rule.

In marketing research, for example, the researchers analyze the past transaction records and then derive some useful information for making proper decisions. An association rule mined from the transaction records may be described as follows.

Rule r1: 88% of the people who buy dried milk also buy beer.

The antecedent and the consequent of rule r_1 are *dried milk* and *beer* respectively, and the confidence of rule r_1 is *88%*. Clearly, with high confidence, rule r_1 is somehow helpful for decision-making.

Wur and Leu [17] proposed an effective Boolean algorithm, named Sparse-Matrix approach (BSM), for mining association rules in large databases. In this approach, two tables, IT and TT, are created to generate *frequent item sets* [17]. Each row in IT_{k-1} represents a frequent item set and each '1' in IT_{k-1} represents an item in item set I. Each '1' in TT_{k-1} represents a record that contains the corresponding item set in IT_{k-1}.

Accordingly, logic OR operation is employed on any two rows in IT_{k-1} to generate a k-item set; meanwhile, logic AND operation is employed on the corresponding rows in TT_{k-1} to generate TT_k. Then, using logic AND and XOR operations, interesting association rules are derived from the frequent item sets in all ITs and TTs.

Consider a simple database with a set I of five items {A, B, C, D and E}, as shown in Table 1. Each row in Table 1 represents a record; for example, row one represents a record (T100) that contains three items (A, D, and E).

Table 1 A simple database

Record number	Items
T100	ADE
T200	BDE
T300	BCDE
T400	AE

By using the BSM approach [17], 14 association rules, as listed in Table 2, are derived from the above database. Notably, each rule i is represented with two indicators: 'Support' and 'Confidence', where 'Support' is the fraction of records that contain the corresponding item sets in both the antecedent and the consequent of rule i; Let W_i denote the records that contain the item sets in the antecedent of rule i. 'Confidence' is the fraction of W_i that contain the item sets in the consequent of rule i.

Table 2 14 association rules obtained from Table 1

Antecedent	Consequent	Support	Confidence
{A}	{E}	50%	100%
{B}	{D}	50%	100%
{B}	{E}	50%	100%
{B}	{DE}	50%	100%
{D}	{B}	50%	67%
{D}	{E}	75%	100%
{D}	{BE}	50%	67%
{E}	{A}	50%	50%
{E}	{B}	50%	50%
{E}	{D}	75%	75%
{E}	{BD}	50%	50%
{BD}	{E}	50%	100%
{BE}	{D}	50%	100%
{DE}	{B}	50%	67%

This work applies the above-mentioned BSM approach for mining association rules from sets of query terms (words). The above logic OR, AND and XOR operations are helpful for the mining tasks in the proposed approach. Accordingly, these association rules are used to create association thesaurus for query expansion. The next section will further detail this idea.

3. Architecture of association thesaurus generation

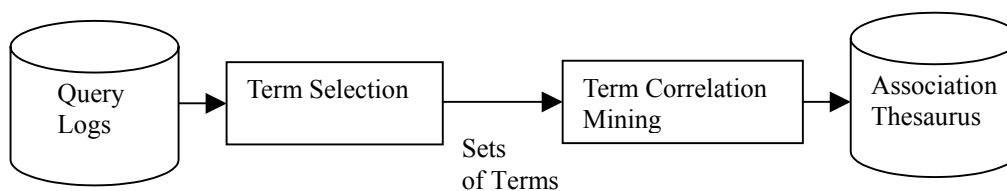This section describes a framework, as shown in Figure 1, to generate association thesaurus for query

expansion. When a user enters a query term, he or she will be presented with search results (for example, web pages or documents). Accordingly, some of these web pages (or documents) are selected by the user. Consequently, the entire information about the entered query term, such as *selected web pages*, user id, and query term, is recorded in the query logs. As mentioned earlier, these records will be used further for association thesaurus generation.

In general, the above selected web pages (or documents) consist of many words (terms). Thus, sets of terms, which are informative in selected web pages (or documents) and useful for the mining tasks, should be identified from selected web pages (or documents). To accomplish this, a value, named Q, is introduced as follows.

$$Q_j = \frac{\sum_{i=1}^{m} tf_{ij}}{m} \times df_j \qquad (2)$$

where Q_j denotes the average term frequency and document frequency of term j, tf_{ij} is the frequency of occurrence of term j in selected page i, df_j is the frequency of occurrence of term j in all selected pages, and m is the total number of selected pages. Clearly, terms with high value of Q(greater than a threshold θ) are informative in selected pages. In this work, these terms are collected as sets of terms and further used for term correlation mining.

Table 3 Five sets of terms extracted from five selected web pages with the initial query term 'neural network'

Selected web pages	Sets of terms
No 1	{training, neuron, learning, unsupervised, weight, perceptron}
No 2	{neural, network, recall, learning, patterns, feature, threshold}
No 3	{learning, training, classifier, patterns, perceptron, layer, weight}
No 4	{network, perceptron, training, patterns, classifier, feedforward, neuron}
No 5	{Hopfield, patterns, weight, optimization, Tank}

Figure 1. Architecture of association thesaurus

Through the above procedure for term selection, various pages 'selected by users' will be transferred into sets of terms (i.e., sets of query terms). For example, as listed in Table 3, five sets of terms may be extracted from five selected web pages respectively, corresponding to the initial query term 'neural network' entered by a user. Accordingly, the above-mentioned BSM algorithm (see Section 2) is applied on these term sets for term correlation mining. Here, different association rules of terms are derived as the following form.

$$r_k: t_i \rightarrow t_j | (s, c), \qquad (3)$$

where t_i and t_j are the antecedent (term) and the consequent (term) of rule r_k, respectively; meanwhile, s and c are the support and the confidence of rule r_k, respectively. Table 4 gives three examples of association rules. Notably, these association rules are asymmetrical [17]; for example, the confidence of rule r_3 (neuron $\rightarrow$ neural) exceeds the confidence of rule r_2 (neural $\rightarrow$ neuron).

Table 4 Three examples of association rules

Rule	Query term	Candidate expansion term	Support	Confidence
r1	neural	network	26.43%	97.34%
r2	neural	neuron	14.91%	27.25%
r3	neuron	neural	11.88%	36.73%

4.Query Agent with association thesaurus

The proposed query expansion method with association thesaurus is implemented within the Query Agent of a course recommendation system, Coursebot [21]. Figure 2 describes the architecture of the Query Agent. To find the desired course pages, a user has to specify a query term via the Interface Agent. Consequently, the Query Expansion module retrieves and ranks candidate expansion terms from the association thesaurus database by using a SQL (Structure Query Language) form (as shown in Table 5); the user is then presented with these expansion terms. Accordingly, the candidate expansion terms selected by the user are combined with the original query for query modification. Finally, the reformulated query is applied again to generate structured course pages for the user.
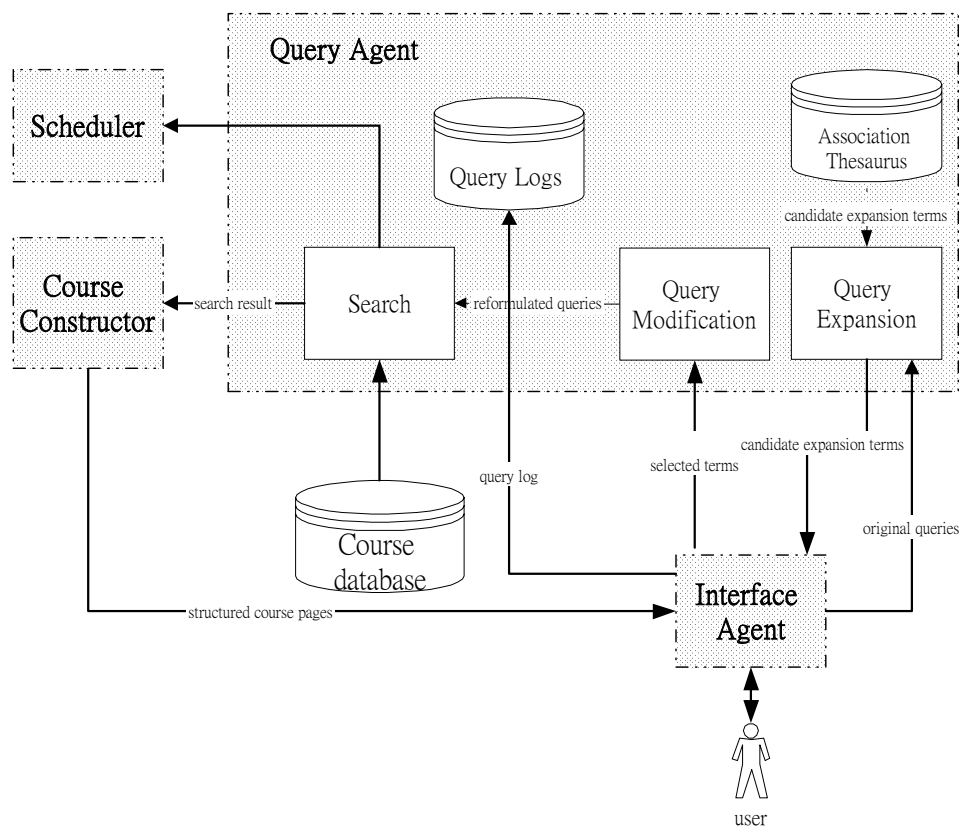

Figure 2. Architecture of the Query Agent

Table 5 SQL form

SELECT * FROM *THESAURUS* WHERE *QUERY = QT* AND *COURSE = C_k* ORDER BY *CONFIDENCE* DESC

where   *THESAURUS* is the database of association thesaurus (rules),
         QUERY is the antecedent of an association rule,
         QT is the original query term,
         COURSE is the course that an association rule belongs to,
         C_k is the id number of course specified by the user, and
         CONFIDENCE is the confidence of an association rule.

5. Experimental Results

As stated above, the proposed query expansion method with association thesaurus is implemented within the Query Agent of Coursebot [21]. Here, 203 query logs of the course 'Neural Network' (The course includes 321 web pages.) in Coursebot was used to demonstrate the performance of the proposed approach. Each query log consists of a query term entered by a user and corresponding selected web pages of the user (The average number of selected web pages in a query is 4.1).
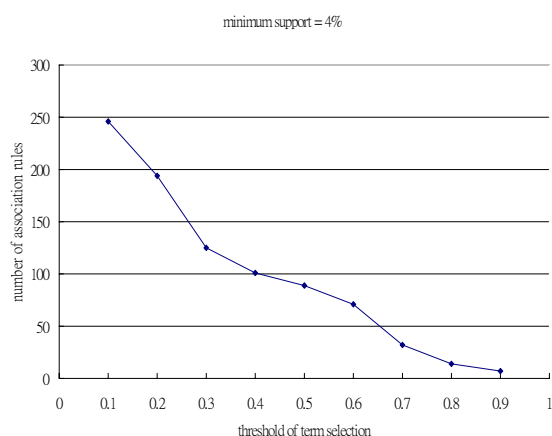

minimum support = 4%

Figure 3. Relationship between the threshold used for term selection and the number of association rules

According to the above-mentioned procedure for term selection (see Section 3), the selected web pages in the above 203 query logs were transferred into sets of terms (i.e., 203 sets of terms). Consequently, various association rules were mined from these 203 term sets. Figure 3 represents the relationship between the number of association rules and the thresholds used for term selection. When querying through the Query Agent of Coursebot, the user will be presented with a thesaurus constructed from the association thesaurus database. For example, Table 6 represents a thesaurus with query term 'learning'

(That is, the antecedent (term) of each association rule in the thesaurus is 'learning').

Table 6 A thesaurus with query term 'learning'

Antecedent	Consequence	Confidence
learning	algorithm	82.78%
learning	network	66.68%
learning	neural	66.25%
learning	function	52.69%
learning	training	50.97%
learning	perception	49.58%
learning	supervised	49.20%
learning	vector	35.41%
learning	output	26.87%
learning	unsupervised	24.30%
learning	system	18.85%
learning	concept	10.24%
learning	backpropagation	8.75%

The proposed approach for interactive query expansion was evaluated by using two ratios, precision ratio and recall ratio, which are defined as follows [13]. (Relevant documents are identified by an expert in "Neural Network")

$$Precision = \frac{Number\ of\ Retrieval\ and\ Relevant\ Documents}{Number\ of\ Total\ Retrieval\ Documents} \quad (4)$$

$$Recall = \frac{Number\ of\ Retrieval\ and\ Relevant\ Documents}{Number\ of\ Total\ Relevant\ Documents} \quad (5)$$

Table 7 compares the precision ratio and the recall ratio when the proposed approach was applied or not. The precision ratio is increased from 61.8% to 91.5%. Also, the recall ratio is increased from 71.7% to 87.9%.

Table 7 Comparison of the precision ratio and the recall ratio

Methods	The proposed approach is not applied	The proposed approach is applied
Precision ratio (%)	61.8	91.5
Recall ratio (%)	71.7	87.9

Moreover, Figure 4 represents the relationship between the above two ratios and the total number of query logs used for mining association rules. It is easily seen that these two ratios can be improved when the total number of query logs used for the mining tasks is increased to a certain degree.

This approach takes advantages of both automatic and interactive query expansion methods because the query is expanded automatically and users are involved implicitly for query expansion. Moreover, the generated association thesauruses offer properties of both similarity thesauruses and synonym-based thesauruses, such as term co-occurrence and synonym-like terms determined implicitly by users.
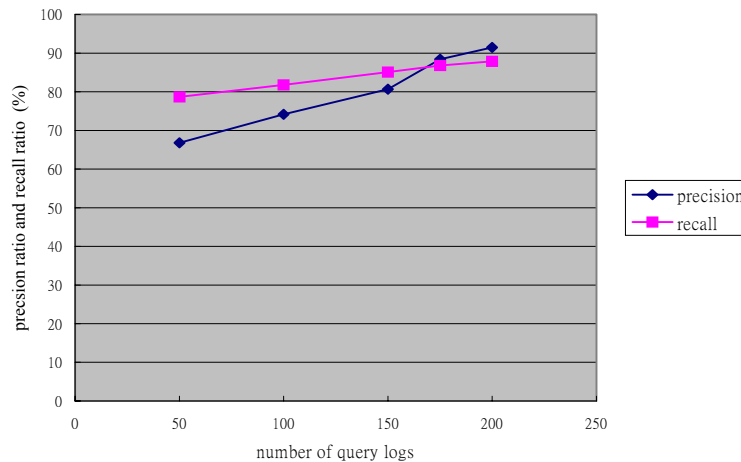
Figure 4. Relationship between the number of query logs used for mining association rules and the two performance ratios of the proposed approach.

6. Conclusions

This paper proposed an interactive query expansion method with association thesaurus, which is mined from the selected web pages of users in the query logs. The selected web pages of users in the query logs are transferred into sets of terms and then used for term correlation mining. Then, various association thesauruses concerning different query terms are constructed from these term correlations. Consequently, the proposed method combines the original query term specified by a user with the corresponding thesaurus to offer the user more precise results.

The query expansion mechanism is implemented within the Query Agent of a course recommendation system, Coursebot. The proposed approach for interactive query expansion was evaluated by using two ratios, precision ratio and recall ratio. Experimental results have shown that the performance of the course recommendation system is improved a lot when the proposed approach is applied.

References

[1] R. Agrawal, T. Imielinki, and A. Swami, "Mining Association Rule Between Sets of Items in Large Database," Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207-216.

[2] E. N. Efthimiadis, "A User-centred Evaluation of Ranking Algorithms for Interactive Query Expansion," Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993, pp. 146-159.

[3] C. Fellbaum, An Electronic Lexical Database, MIT Press, Cambridge, Mass., 1998.

[4] B. Jansen and A. Spink, "Methodological Approach in Discovering User Search Patterns Through Web Log Analysis," Bulletin of the American Society for Information Science and Technology, 27(1), 2000, pp. 15-17.

[5] K. Jarvelin, J. Kristensen, T. Niemi, E. Sormunen, and H. Keskustalo, "A Deductive Data Model for Query Expansion," Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996, pp. 235-243.

[6] W. S. Li, and D. Agrawal, "Supporting Web Query Expansion Efficiently Using Multi-Granularity Indexing and Query Processing," Data & Knowledge Engineering, 35, 2000, pp. 239-257.

[7] M. Magennis and C.J. van Rijsbergen, "The Potential and Actual Effectiveness of Interactive Query Expansion," Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in information Retrieval, 1997, pp.324-332.

[8] J. McDonald, W. Ogden, and P. Foltz, "Interactive Information Retrieval using Term Relationship Networks," Proc. of the 6th Text Retrieval Conference, 1997.

[9] G. A. Miller, "WordNet: A Lexical Database," Comm. ACM, 38(11), 1993, pp. 39-41.

[10] Y. Qiu and H. P. Frei, "Applying a Similarity Thesaurus to a Large Collection for Information Retrieval," Technical Report. Dept. Computer Science, Swiss Federal Institute of Technology (ETH), Jan 1995.

[11] Y. Qiu and H. P. Frei, "Concept Based Query Expansion," Proc. of the 16th ACM SIGIR International Conference on Research and Development in Information Retrieval, 1993, pp. 160-169.

[12] J. J. Rocchio, "Relevance feedback in information retrieval," In the SMART Retrieval System-- Experiments in Automatic Document Processing, pp. 313-323, Englewood Cliffs, NJ, 1971. Prentice Hall, Inc.

[13] G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill, 1983.

[14] B. R. Schatz, E. H. Johnson, and P. A. Cochrane, "Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval," Proc. of the 1st ACM International Conference on Digital Libraries, 1996, pp. 126-133.

[15] F. Sebastiani, "Automated Generation of Category-Specific Thesauri for Interactive Query Expansion," Proc. of the 9th International Databases Conference on Heterogeneous and Internet Databases, 1999, pp. 429-432.

[16] A. Spink, "Term Relevance Feedback and Query Expansion: Relation to Design," Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in information Retrieval, 1994, pp. 81-90.

[17] S. Y. Wur and Y. Leu, "An Effective Boolean Algorithm for Mining Association Rule in Large Databases," Proc. of the 6th International Conference on Database Systems for Advanced Applications, 1998.

[18] Altavista, available at http://www.altavista.com

[19] Excite, available at http://www.excite.com

[20] Google, available at http://www.google.com

[21] The Coursebot, available at http://coursebot.et.ntust.edu.tw

[22] The Emerging Digital Economy, available at http://www.ecommerce.gov