Ontology-Supported Web Recommender for Scholar Information

Sheng-Yuan Yang\textsuperscript{1} Chun-Liang Hsu\textsuperscript{2}
\textsuperscript{1}Department of Computer and Communication Engineering, St. John’s University
\textsuperscript{2}Department of Electrical Engineering, St. John’s University
E-mail: \{\textsuperscript{1}ysy, \textsuperscript{2}liang\}@mail.sju.edu.tw

Abstract

In this quickly developed and shifting era of Internet, how to make use of webpage indexing structure or search engines which let information demanders fast and precisely search and extract out advantage information has become extremely important capability in users on the Web. This paper combined a data mining tool SPSS Clementine with the domain ontology to mine out useful important information from huge datum, and then to employ Java to develop an information recommender for scholars--- OntoRecommender, in which can recommend suitably important information to scholars. The preliminary experiment outcomes proved the reliability and validation of the recommender achieving the regular-level outcomes of information recommendation, and accordingly proved the feasibility of the related techniques proposed in this paper.

1. Introduction

Along with popularity of application and use of Internet, search tools and their related techniques have become a necessary information accomplishment in daily life. In this information-exploding era, there is not only a big trouble in the information quantity, but also an important issue on how to get the suitable information. Google was officially born and properly provided the information query mechanism to look after both sides of quality and quantity: keyword query. However, this query method resulted in a hard-breathing searching outcome even with less keyword. In such long and complicatedly listed outcomes, it caused users not only took more time to look up those information but also indicated the query demands of users. The main cause was that the keywords entered by users were not completed and not able to obviously indicate the query demands of users. Furthermore, there are so many keywords being the same words with different meaning in different fields. The system would finally produce many complicated cross-field query outcomes when system didn’t respectively managed to classifying query requisition and specifying fields. The appearance of ontology is exactly a powerful tool for solving problems described before.

Data mining employed the techniques of statistic analysis, information classification, or machine learning to provide important decision bases of information systems in according with significant reasons, relationships, or their potential rule models found from the huge amount of datum. It usually includes two basic models: one is from up to down to verify whether thinking was right (hypothesis checking), it can employ the techniques of mathematical statistics to do the hypothesis checking; the other is from down to up to mine its potential facts from the huge amount of datum (knowledge discovery), it can employ the information analysis techniques of statistic methods or artificial neural networks to do that. The six common data mining methods include description, estimation, clustering, classification, prediction, and association [9]. There exist properly related techniques and algorithms for each method. In this paper, we combined classification and related association rules to do information mining and accordingly provided the significant bases to the backend system for further processing.

To do a good data mining need to have an exact and complete process of data mining. CRISP-DM (short for CRoss-Industry Standard Process for Data Mining) is a generally acknowledged standard process of data mining, which was divided into six stages: business understanding, data understanding, data preparing, modeling, evaluation, and deployment. In the whole process, it makes information demanders orderly and precisely proceeding with data mining in each stage. Hence, whether there exists a data mining tool with ease to learn and strong functions to accordingly make the mining process effectively to get yield twice the result with half the effort. SPSS Clementine is an exactly effective tool with above mining process and functions [9].

To sum up, the topic of this paper was to employ the ontology technique to design the ontology of important information of some scholars, and then accompanied with MS SQL Server database to set up an ontology sharing platform of some keywords of some scholars’ information and applied with SPSS Clementine to be a mining tool of huge amount of information classification and recommending implementation. Finally, we use Java [7] to build up the OntoRecommender (short for Ontology-supported Recommender). In other words, introducing classification comparison and association analysis of scholars’ information ontology not only exclude the error problems of information analysis producing from the man-made subjective factor but also go pliable but strong the association analysis of information.
classification as well as reach support for both fast and effective recommender system of scholars’ significant information.

2. Related Background and Developing Techniques

2.1. Ontology

Ontology [3] was one theory in philosophy and primarily to explore knowledge characteristics of life and real objects; in artificial intelligent field it was used to define the content of domain knowledge, express knowledge, solve communication, and commonly share problems; in information technology field it offered much assistant for research and development of E-commerce and Knowledge Management [14,15]. Ontology provides complete semantic models, which means in specified domain all related entities, attributes and base knowledge among entities have sharing and reuse characteristics which could used solving the problems of common sharing and communication. To describe the structure of the knowledge content through ontology can accomplish the knowledge core in a specified domain and automatically learn related information, communication, accessing and even induce new knowledge; hence, ontology is a powerful tool to construct and maintain an information system [14].

2.2. Related Mining Tools and Developing Techniques

Data mining adapted the common statistic software of computer science domain SPSS as a basic tool, which could save complicated processes of mathematic derivation but focus on statistic applications in computer science. Clementine is one of the software series of SPSS which was developed by the standard CRISP-DM for data mining. It adapted the unique work stream to proceed with data mining, which can make information recommenders easily, fast, and effectively accomplish data mining works. Fig. 1 illustrates the tool interface of SPSS Clementine (Chinese version). It offers various basic functions data work stream in the bottom selection plate, and accordingly provides related data mining treatments for building up work stream in the middle left block.

This system employed Java to develop the front-end system interface and adapted MS SQL Server as backend knowledge-database sharing platform based on ontology. MS SQL Server is one mostly used relational database management system. SQL (Structured Query Language) is one common query language of relational database to get the data in the database.

3. System Structure and Process

3.1. Construction of Ontology Database

The source of the ontology database in this paper derived from all previous sessions’ presidents, supervisors, directors of Taiwanese Association for Artificial Intelligence and Taiwan Fuzzy Systems Association and related professors in Artificial Intelligence (AI) domain of universities and colleges in Taiwan, which fetch out the professional keywords and their synonym appearing in their giving lessons and research areas, respectively, were used as the construction bases of the ontology database in this paper. The keyword weight equals the number of keyword appearances in webpage of each domain professor. For example, webpages of ten professors in Artificial Intelligence domain appear nine times of terms “Artificial Intelligence” and the system set the term weight to 0.9; and the term embedded systems appeared six times and its weight was set to 0.6; the rest of terms weights are the same as above, detailed part of the ontology database shown in Fig. 2.

3.2. System Operation and Structure

Fig. 3 showed the operation system structure of OntoRecommender, and related techniques and functions of every part were described as below.

(1) CKIP segmentation system: the CKIP (Chinese Knowledge and Information Processing) group is a research team formed by the Institute of Information Science and the Institute of
Linguistics of Academia Sinica in 1986. Its purpose is to establish a fundamental research environment for Chinese natural language processing. The preliminary goal of the project was to construct research infrastructures with reusable resources that could be shared by domestic and international research institutes. This research employed the CKIP segmentation system to be a front-end assistant tool, which segment webpage contents of scholars and filter out most part of stop words. Its output format is a text file.

Scholar webpage
CKIP Segmentation System
Segmentation Fixing
Stop Word Filtering
JAVA System

Ontology Database
SPSS Clementine System
Term frequency statistics
Recommendation Mechanism
Keyword Classifying
Best Recommending
Output as Text File

Fig. 3 Operation system structure of OntoRecommender

(2) Segmentation Fixing: the CKIP segmentation system could make segmentation errors in some terminologies of specific domains, for example, 「臺灣科技大學」(National Taiwan University of Science and Technology) could be segmented to 「臺灣」(Taiwan), 「科技」(Science and Technology), 「大學」(University). Those errors could make an enormous wrong effect in the precision rate of backend word matching. The function of this block emphasized to solve those problems, detailed function descriptions as below [16]:

(i) Preprocessing: load stop word database StopWords.txt, which currently contains about 1700 stop words; delete those data in webpages having nothing to do with classification such as Tab, \n, punctuation marks, and continuous space; finally, store input document into string array SourceText so as to manage afterwards.

(ii) Segmentation: divide SourceText into the least semantic unit - vocabulary. Searching every space character in the document from the first character until the next one, and this is what we defined single phrase. Then system extracted every single phrase and stored into array [12].

(iii) Stop word filtering: use “stop list” to store those words and ought to be excluded when indexing words and phrases so as to decrease the noise in document and increase classification precision.

(iv) Term frequency statistics: build up one terms-frequency array corresponding to vocabulary array from the first vocabulary as the base. After comparing with every vocabulary, if one vocabulary appeared repeatedly, then we add one recode to the tem-frequency. Then delete the repeated term and tem-frequency record till the last vocabulary. Finally, calculate number of appearances of each vocabulary used as the classification bases.

(3) SPSS Clementine System: enter the text file after processed by Segmentation Fixing and combine against the domain ontology database to be the bases of classification statistics and recommending match analysis. This processing divided into two stages: the first stage judged the scholar webpage whether fit in with specific domain, i.e., content-based filtering; the second stage extracted the significant information from the scholar webpage to be the important recommending base. In the recommending mechanism, not only has the general normality recommending but also adding the best recommending module. For example, if existed the duplication list of giving lessons between a scholar’s and some related scholars’, these significant courses will be the best information of “Courses” recommending, i.e., collaborative filtering.

(4) Recommender Display: show the recommending results after mining by SPSS Clementine. This system employed Java to implement the user interface of this recommender.

4. System Display and Verification

4.1. System Display

Fig. 4 Scholar’s webpage content after segmented by CKIP segmentation system

Fig. 4 showed a scholar’s webpage content after segmented by CKIP segmentation system. The left block of Fig. 5 (i.e., Segmentation Area) illustrated the result of the former after deleted stop words and calculated term-frequency of related vocabularies. Fig. 6 showed the screen of related mining stream...
constructed with SPSS Clementine. The right block of Fig. 7 (i.e., Recommending Area) illustrated the significant scholar information mined by SPSS Clementine via this recommender. The significant recommending information of scholar divided into Courses, Academic Activities, and Related Introduction in according to the best recommending, AI, Fuzzy, and NN (Neural Network).

4.2. System Verification Rules

The recommending information meant the best recommending have chosen from a group of related information sets. That wonderfully possessed different approaches to the same purpose as whether sampling specimens can be on behaving of degree of sampling body in huge amount of datum. In the sampling survey domain, the reliability was usually employed to measure the degree of precision of sampling system itself, while the validity was emphasized whether it can be correct to reflect the properties of the appearance of things [13]. In other words, the former evaluates the stability of the measurement tool, while the latter focuses on the correctness of the tool itself.

In 1979, J.P. Peter [8] had the aid of mathematic model to represent the definitions of the reliability and validity, detailed as below.

(1) Reliability

To assume a measurement tool measured the value $X_o$ (generally is assumed to be the mean value), which can be divided into:

$$X_o = X_t + X_e$$  (1)

where, $X_o$ means observed $X$, $X_t$ means true $X$, and $X_e$ means error $X$. The variance of the measured/observed value also is assumed to be the $V_o$, which can be also divided into:

$$V_o = V_t + V_e$$  (2)

where, $V_o$ means observed $V$, $V_t$ means true $V$, $V_e$ means error $V$. A reliability coefficient ($r_{tt}$), therefore, is nothing more than to the ratio of true variance to observed variance:

$$r_{tt} = \frac{V_t}{V_o}$$  (3)

Because $V_t$ cannot be estimated directly from statistic view, therefore, equation (3) can be rewritten into a computational formula as:

$$r_{tt} = \frac{(V_o - V_e)}{V_o} = 1 - \left(\frac{V_e}{V_o}\right)$$  (4)

In other words, the reliability equals 1 minus the ratio of error variance to observed variance.

(2) Validity

If $V_t$ can be divided into $V_{co}$ plus $V_{sp}$ again, then

$$V_o = V_{co} + V_{sp} + V_e$$  (5)

where, $V_{co}$ means correlated $V$ which is the common variance related to measurement properties, $V_{sp}$ means specific $V$ which is the individual variance unrelated to measurement properties. The definition of validity $V_{al}$ is:

$$V_{al} = \frac{V_{co}}{V_o}$$  (6)

(3) Mathematic relationship between the reliability and validity:

Let $V_o = V_{co} + V_{sp}$ substitute for equation (3) to get

$$r_{tt} = \frac{V_t}{V_o} = \frac{(V_{co} + V_{sp})}{V_o}$$
$$= \frac{V_{co}}{V_o} + \frac{V_{sp}}{V_o}$$  (substitute for equation (6))
$$= r_{tt} - V_{al}$$

In other word, $V_{al} = r_{tt} - V_{al}$, i.e., the validity $V_{al}$ should be not greater than the reliability $r_{tt}$.

4.3. Data Verification

The recommending significant information of this experiment was asserted by the domain experts,
including observed values, true values, error values, and related variances. Table 1 were used equation (4) to calculate the reliabilities of some scholars’ recommending information on “Courses” and “Academic Activities” which are 0.856 and 0.756 in average, respectively. Table 2 were used equation (6) to calculate the validities of some scholars’ recommending information on “Courses” and “Academic Activities” which are 0.856 and 0.756 in average, respectively. Finally, what merits attention is: the Professional Classification of each scholar can be accurately shown that prove the recommending structure we proposed in this paper has its accuracy and availability.

Table 1 Results of the reliability of classification information

<table>
<thead>
<tr>
<th>Scholar</th>
<th>Courses</th>
<th>Academic Activities</th>
<th>Professional Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.S. Ho</td>
<td>V₁v₂ 0.92</td>
<td>V₁v₂ 0.71</td>
<td>Al</td>
</tr>
<tr>
<td>T.W. Kao</td>
<td>V₁v₂ 0.21</td>
<td>V₁v₂ 0.71</td>
<td>Al</td>
</tr>
<tr>
<td>S.Y. Yang</td>
<td>V₁v₂ 0.51</td>
<td>V₁v₂ 0.71</td>
<td>Al</td>
</tr>
<tr>
<td>S.M. Chen</td>
<td>V₁v₂ 0.36</td>
<td>V₁v₂ 0.57</td>
<td>Fuzzy</td>
</tr>
<tr>
<td>W.L. Hsu</td>
<td>V₁v₂ 0.57</td>
<td>V₁v₂ 0.5</td>
<td>Al</td>
</tr>
<tr>
<td>Average</td>
<td>0.856</td>
<td>0.756</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Results of the validity of classification information

<table>
<thead>
<tr>
<th>Scholar</th>
<th>Courses</th>
<th>Academic Activities</th>
<th>Professional Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.S. Ho</td>
<td>V₁v₂ 0.92</td>
<td>V₁v₂ 0.71</td>
<td>Al</td>
</tr>
<tr>
<td>T.W. Kao</td>
<td>V₁v₂ 0.21</td>
<td>V₁v₂ 0.71</td>
<td>Al</td>
</tr>
<tr>
<td>S.Y. Yang</td>
<td>V₁v₂ 0.51</td>
<td>V₁v₂ 0.71</td>
<td>Al</td>
</tr>
<tr>
<td>S.M. Chen</td>
<td>V₁v₂ 0.36</td>
<td>V₁v₂ 0.57</td>
<td>Fuzzy</td>
</tr>
<tr>
<td>W.L. Hsu</td>
<td>V₁v₂ 0.57</td>
<td>V₁v₂ 0.5</td>
<td>Al</td>
</tr>
<tr>
<td>Average</td>
<td>0.856</td>
<td>0.756</td>
<td></td>
</tr>
</tbody>
</table>

5. Related Works and Comparison

With the increasing applications of recommendation systems in a variety of fields, various approaches to recommender systems have been developed. There are three paradigms for recommendation systems, namely, content-based filtering, collaborative filtering, and hybrid filtering approaches [5]. Content-based recommender systems base their recommendations on the similarity between new items and ones that users liked before, which described items with metadata and extracted features. An example of content-based recommender systems is: Chang et al [4] proposed an approach trains the artificial neural networks to group users into different clusters and applied the well-established Kano’s method to extracting the implicit needs from users in different clusters for the improvement of information overloading in a real case of tour and travel websites. Collaborative recommender systems generate their recommendations obtained from people with similar interests, i.e., people with the same interests may tend to exhibit the same behavior and require the same information. Here is an example of such recommender systems. Liang et al [6] explored how to utilize tagging information to do personalized recommendation. The content-based and collaborative filtering approaches are not mutually exclusive to each other, and there have been many efforts to integrate them in order to obtain more accurate recommendations, i.e., hybrid recommender systems, which take the advantages of both features in a single framework. The feedback required for content-based recommendation is shared, allowing collaborative recommendation as well. There is an example of hybrid recommender systems. Zhang and Jiao [17] proposed an associative classification recommendation system for personalization in B2C e-Commerce application.

Ontology [3] was one theory in philosophy and primarily to explore knowledge characteristics of life and real objects; in recommender system field it can play in the adoption of ontologies for modeling the domain for contributing to tailor the right information/service to users and thus facilitate the user-system interaction and the system communication with other agents [1]. As ontology techniques become more and more mature, lots of researchers have begun to explore their applications to recommendation systems. Cantador et al [2] presented New@hand, a news recommender system which makes use of semantic technologies to provide personalized and context-aware recommendation. Weng and Chang [10] proposed to use ontology and the spreading activation model in research paper recommendation for elevating the system performance and also improving the shortcomings of today's recommendation systems. In this paper, an ontology-supported information integration and hybrid recommender system for scholars was proposed, not only can it solve the problem of so many keywords being the same words with different meaning in different fields which resulting in many complicated cross-field recommendations, but also it can fast integrate specific domain documents and extract significant information from them to take information integration and recommendation ranking. In addition, we reckon that the data mining techniques are more and more growing and mature. Its analysis and induction features, supported by the underlying online/offline mining capability, however, provides yet another level of automation in recommendations and deserves more
attention, such as using the online analytical processing (OLAP) ability of data warehousing to solve the contradicting problems among hierarchy rating mentioned in [11]. That is a main reason of why we combined a data mining tool SPSS Clementine with the domain ontology to mine out usefully important information from huge datum.

6. Conclusion

This paper has applied the ontology technique to design the ontology of some scholars’ significant information, and then accompanied with MS SQL Server database to set up an ontology sharing platform of some keywords of some scholars’ information and applied with SPSS Clementine to be a mining tool of huge amount of information classification and recommending implementation. Finally, we use Java to build up the OntoRecommender. In other words, introducing classification comparison and association analysis of scholars’ information ontology not only exclude the error problems of information analysis producing from the man-made subjective factor but also go pliable but strong the association analysis of information classification as well as reach support for both fast and effective recommender system of scholars’ significant information. The preliminary experiment outcomes proved the reliability and validation of the recommender achieving the regular-level outcomes of information recommendation, and accordingly proved the feasibility of the related techniques proposed in this paper.

This OntoRecommender could easily combine with backend systems due to both an open-sourced code design philosophy and the more developing tool SPSS Clementine adapted by enterprises so as to attract related researchers to use this technique. Continuously improving the performance efficiency, expanding database of ontology and its related linking interface, and developing the middle programs with data mining tools would be the everlasting research in the future.

Acknowledgement

The authors would like to thank Sie-Yi Luo and You-Ren Jhang for their assistance in system implementation and experiments. In the meantime, the authors also specially thank the department of Computer and Communication Engineering in St. John’s University in support on related development equipments and environment.

References


