Exploring Business Ecological Information with Personal Attention

Fu-Ren Lin  
National Tsing Hua University, frlin@mx.nthu.edu.tw

An-Yun Lee  
National Tsing Hua University, amber21.lee@gmail.com

Ching-Pin Cheng  
National Tsing Hua University, g9773520@oz.nthu.edu.tw

Follow this and additional works at: http://aisel.aisnet.org/pacis2013

Recommended Citation
http://aisel.aisnet.org/pacis2013/153

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2013 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
EXPLORING BUSINESS ECOLOGICAL INFORMATION WITH PERSONAL ATTENTION

Fu-ren Lin, Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan, R.O.C., frlin@mx.nthu.edu.tw
An-yun Lee, Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan, R.O.C., amber21.lee@gmail.com
Ching-pin Cheng, Institute of Technology Management, National Tsing Hua University, Hsinchu, Taiwan, R.O.C., g9773520@oz.nthu.edu.tw

Abstract

In the era of Web 2.0, the amount of information has been growing exponentially, intensifying the information overload problem. The inconvenience caused by information overload not only lies in casual search by every Internet user, but also in finding business related information. Business analysts now need to explore an overwhelming amount of information in order to understand business ecosystem or technology development. This study designed a mechanism hoping to decrease the information searching cost and improve the effectiveness of information searching as well. Based on the business ecosystem identified by relation extraction technique, we combined semantically networked knowledge base to organize information of diverse sources to facilitate exploratory search. On top of that, we adopted personalization technique to filter irrelevant information, and applied spreading activation model in the business ecological knowledge. Through experimental design and evaluation, we found that users can identify more relevant information facilitated by the built system than searching only via term matching or user profile matching.

Keywords: business ecological information, information retrieval, personal attention, knowledge map.
1 INTRODUCTION

1.1 Research background

The rapid growth of Internet contents with search facilitation makes information easy for users to acquire and disseminate, which also influences the professionals in adding value on top of the information retrieved via the Internet. Moreover, in Web 2.0 era, users are also the information contributors, and significantly increase internet contents; for example, 487 billion GB of information created in 2008 (Gantz, Boyd, and Dowling, 2009). IDC research suggested that the global data load would rise from 0.8zb in 2009 to 35zb in 2020, and even this might understate the case. The improvement of technology and the trend of sharing information had been making the task of retrieving documents of interest easier than before; however, it takes much more effort on comprehending the contents retrieved due to the problem of information overload. According to Basex, there is 650 billion US dollar value of productivity lost per year as a result of information overload. One symptom of information overload is the inability to get the right information to workers at the right time (Gantz, Boyd, and Dowling, 2009).

The availability of daily updated contents is of great value without doubt, especially for volatile information for emergent technologies; it intensifies the information overload problem. It becomes even more difficult for individuals to identify interesting information from many daily updated content sources, such as blog entries, news and scientific publications (Nanas et al., 2010). The inconvenience caused by information overload not only occurred in casual search by general internet users, but also in finding business related information. Business analysts now need to explore an overwhelming amount of information in order to understand business ecosystem or technology development. Marchionini (2006) described exploratory search as the activity of finding and understanding knowledge about a topic of interest (Castano et al., 2011). According to Marchionini (2006), exploratory search includes search activities with the purposes of leaning and investigation. For learning objective, users seek information to acquire knowledge and to comprehend concepts. Investigative searching is for planning and forecasting, or to transfer data into knowledge (Marchionini, 2006). For users, such as intelligence analysts, they perform exploratory searches every day as a part of their job (Ahn et al., 2010).

With individual’s constraints, such as limited time, energy and skills, the amount of information we can process is limited. At the same time, much of business information is text subject to frequent changes. With the progress of electronic commerce and pervasiveness of information-intensive business operations, the ability to retrieve information effectively is of economic importance (Gao et al., 2005). Hence, the use of efficient and effective mechanisms to retrieve required business information is crucial to business analysis, and that saves users time and energy to concentrate on value-added works, such as information interpretation and analysis. In addition, with more relevant information available also reduces uncertainty and helps decision makers make rational decisions (Citroen, 2011). Therefore, in order to reduce the information searching and processing costs for users, e.g., business analysts, we aim to design an information retrieval system, which is capable of automatically retrieving and recognizing useful information.

2 LITERATURE REVIEW

Enterprises do business analysis every day in order to stay competitive. James F. Moore defined business ecosystem as “an economic community supported by a foundation of interacting

1 http://www.v3.co.uk/v3-uk/news/1990499/information-overload-exceed-predictions
2 www.basex.com
organizations and individuals--the organisms of the business world.” The economic community produces goods and services of value to customers, who themselves are members of the ecosystem.

However, with exponentially growing amount of information, business ecosystems have become more complicated than ever. Analysts need to spend more effort on examining business ecosystems. Thus, the research combines personalization technique, semantic web and spreading activation theory to help increase the effectiveness of information search.

2.1 Relation extraction

Cheng (2010) developed a technology monitoring service for business ecosystem. The ecosystem is identified by name entity recognition and relation extraction from online news articles⁴. The system used OpenCalais⁵ to recognize name entities from online news articles, and extracted company, product and technology, three types of name entities from various news sources. Among all services in the framework, the relation extraction is the most important one (Cheng, 2010).

Since the main focus of the research lies in discovering business ecosystem relationship consisting of company, product and technology, Cheng’s research extracted six types of name pairs, which are (1) company-company, (2) technology-technology, (3) product-product, (4) company-technology, (5) company-product, and (6) technology-product (Cheng, 2010). After name entities are extracted, relations between them are also identified by relation extraction technique modified from RelEx (Fundel et al., 2007). An information filtering mechanism is designed based on the business ecosystem identified from the technology monitoring system.

2.2 Personalization

While hoping to find more information, there is a tradeoff users must make between searching through fewer articles and finding more information of personal interest (Lang, 1995). One way to assist a user in dealing with the task of finding relevant information is through the personalization of web content, for example, the customization of a web page’s content according to an individual’s interests (Nanas et al., 2010). Personalization is defined as “the ability to provide content and services tailored to individuals based on knowledge about their preferences and behavior,” or “the use of technology and customer information to tailor electronic commerce interactions between a business and each individual customer” (Adomavicius & Tuzhilin, 2005). It is a process of gathering and analyzing user information to deliver proper information (Kuo & Chen, 2001).

Liang et al. (2007) introduced two types of recommendation mechanisms: attribute-based and collaborative filtering. The former one uses product features as attributes, and could be used for content-based filtering. Collaborative filtering uses user behavior as the filtering attribute. Content-based filtering performs better for items with descriptive features that can be automatically extracted, because it builds user profile according to a user’s browse history. The user profile would thus be composed of features selected from browsed items. The method provides recommendation for new-coming items based on user profile, so that it doesn’t require a rating matrix and can easily deal with the short life span of news stories (Nanas et al., 2010). The advantages for content-based filtering are user independence, transparency with new item (de Gemmis et al., 2009). For these reasons, content-based filtering is better suited to news personalization than collaborative filtering (Billsus & Pazzani, 2007).

---

⁴ The news sources are: Business Week, Fast Company, Tech Crunch, Engadget, NYTimes, Venture Beat, Wired, ZDnet, DGTimes and RedHerring
⁵ http://opencalais.com
2.3 Semantic web

Many researches applied abstract concepts to organizing and connecting related information, to facilitating information searching and discovery. Castano et al. (2011) recognized entities from tagged resources, microdata resources, and semantic web resources on the web, and used a unified form of representation to combine these resources into an information cloud. Therefore, a knowledge base can be used for integrating independent information on the web, organizing it into a more useful manner. Past researches (Ochs et al., 2011; Passant & Raimond, 2008; Stan et al., 2011) utilized semantic web to support exploratory search. A well-known one is the Linking Open Data (LOD) project6 proposed by Berners-Lee. This project connected related data that weren't previously linked on the Internet, such as wikipedia, FOAF, wordNet, and Myspace, using URIs and RDF to form a large knowledge base.

Semantic network is a common data structure used in knowledge bases. Since introduced by Quillian (Quillian, 1968), semantic networks have played a significant role in knowledge representation research. According to Quillian’s definition, semantic networks express knowledge in terms of concepts, their properties, and the hierarchical sub-superclass relationship between concepts. Each concept is represented by a node and the hierarchical relationship between concepts is denoted by connecting appropriate concept nodes via “is-a” or “instance-of” links (Schiel, 1989).

Therefore, we can connect relevant information through semantic web to assist users in finding knowledge related to the search target. Applying the semantic web on the Internet, such as Linked Open Data, into our news repository can thus help users discover relations between name entities which were not mentioned in the same article.

2.4 Spreading activation theory

The Spreading Activation (SA) Theory was developed based on the way people process memory (Crestani, 1997). It can be used as a mechanism to interpret how semantic network functions in human brains in cognitive psychology (Collins & Loftus, 1975). Based on the theoretical foundation that human’s left hemisphere uses semantic association structure to process information (Posner & Petersen, 1989), we applied SA Theory to processing information in the semantic network. In the information retrieval (IR) field, lots of works related to processing information in the semantic network, making SA Theory widely used in IR research, and one of the applications is associative retrieval.

SA Model based on SA Theory is a processing technique for network data structure (Crestani, 1997). The processing technique starts from a set of nodes, each with an initial activation value, and after an iteration of activation, the value of a node propagates to its neighboring nodes (Jiang & Tan, 2009). Nodes with propagated values are activated. After each iteration, a termination condition is checked to determine whether the process continues. The termination condition could be a threshold for activation value or distance constraint, where the process ceases when it reaches nodes too far from the initially activated nodes (Crestani, 1997). The process can be represented by $V_i = \sum_{j=1}^{n} (O_j * R_{ij})$, where $V_i$ is the total input to node $i$. $O_j$ is the output from node $i$ to node $j$, and $R_{ij}$ is a weight associated to the link connecting node $i$ to node $j$ (Crestani, 1997). As SA Model has been proven to be efficient for inferencing semantic networks (Jiang & Tan, 2009), we adopted the mechanism to find related information in the business ecosystem.

---

6 http://linkeddata.org/
3 SYSTEM FRAMEWORK

Hevner et al. (2004) proposed the design science framework by which IS research could be built on rigorous past research, and be used for creating value by solving real-world problems. This research used personalization and spreading activation mechanism as the technical foundation to facilitate information filtering, and combined Linked Open Data to link name entities retrieved from news articles, building semantic data network to explore information in the semantic web. Given that a user’s attention could be recorded by his/her clicks on articles they read, the system could keep track of the user’s attention and create personal knowledge map based on these clicked articles. By matching the knowledge map and content items, the information retrieval system can help users discover relevant knowledge.

3.1 System architecture

The system consists of two major modules: user browsing and knowledge map maintenance, as shown in Figure 1. Once a user sends a query, the system first does query expansion through spreading activation, retrieve a set of relevant news articles, and rank articles by the user’s personal attention.

3.2 User browsing module

Cheng (2010) utilized OpenCalais to extract name entities (NEs) (including company, product, and technology) and relations between them from news articles, formed a business ecological network by news data. Based on the network, we used spreading activation (SA) as the query expansion mechanism to search related name entities for the initial search terms. The input value of node $i$ ($V_i$) can be calculated by $V_i = \sum_{j=1}^{n} (O_j * R_{ij})$, where $O_j$ is the output value of node $j$ in an iteration, $R_{ij}$ is the weight of relation from node $j$ to node $i$, and $n$ is the number of nodes linked to node $i$. After an iteration of spreading, node $i$ sends its value to its neighboring nodes, with an output value $O_i$, derived from $O_i = V_i * W_i$, where $W_i$ denotes the weight for node $i$, $V_i$ is the spreading value from the latest iteration. A set of relevant keywords will be collected after query expansion. The system then retrieves news articles, which are the initial retrievals, containing these keywords. Each article is given an attention score which is the sum of attention scores of these NEs facilitated by knowledge maintenance module, and the retrieved articles are ranked decendently according to their attention scores.

![Figure 1. System framework](image-url)
3.3 Knowledge map maintenance module

A knowledge map is composed of networks of nodes, where a node denotes a name entity (NE), and a link between two entities denotes the relation between NEs. The knowledge map is formed by linking nodes semantically with relation extracted from news articles, assigning a personalized weight at the same time. Two components are responsible for maintaining the knowledge map elaborated as following subsections.

3.3.1 Personal attention tracking

For an individual user’s knowledge map, each NE is marked by a personal attention score and a popularity score. An NE is popular while many news readers concern hot events of the day (Lavie et al., 2010). We combine these two measurements to represent a user’s reading behavior. Thus, a score for NEi is calculated by \( W_i = w_a * A_i + w_p * P_i \), where \( A_i \) is the personal attention score to \( NE_i \), \( P_i \) is the popularity score of \( NE_i \), and \( w_a \) and \( w_p \) are the weights for both scores, respectively.

The personal attention score to \( NE_i \), \( (A_i) \) is derived from the interaction between a user and the system. In this study, a user’s click of an article denotes the user’s interaction on the article. \( A_i \) is calculated as the total clicks of \( NE_i \) divided by the sum of clicks of all NEs. \( w_a \) is calculated as the number of clicks on NEs with attention divided by the sum of clicks of all NEs. The popularity score \( (P_i) \) of NE is calculated as the number of articles containing \( NE_i \) divided by the total number of articles. \( w_p \) is calculated as the number of clicks on popular NEs divided by the total number of articles.

3.3.2 Semantic relation tracking

In coding the relations between two NEs, we integrate extracted relations from news articles and semantic relations specified in Wikipedia, considering the dynamic and static relationships respectively to represent semantic relations between NEs. The relation score \( (R_{ij}) \) between \( NE_i \) and \( NE_j \) is calculated as \( R_{ij} = w_s * S_{ij} + w_r * T_{ij} \), where \( S_{ij} \) stands for the relation extracted from news article, \( T_{ij} \) stands for the relation from ontological concepts, and \( w_s \) and \( w_r \) are the weights for both relation scores, respectively\(^7\).

Since the NEs in a knowledge map are extracted from news articles, we utilize relations mentioned in the news to represent the link between NEs. The relation score \( (S_{ij}) \) is calculated as the number of relations between \( NE_i \) and \( NE_j \) divided by the total number of relations \( NE_i \) has.

Apart from the relations extracted from news articles, we also used ontology information to tag categories to name entities. Since a large amount of semantic information may exist in most documents, it is useful to include this information in making recommendations (Liang et al., 2008). We chose DBpedia as the knowledge base used for tagging because the DBpedia project extracts structured information from Wikipedia and converts into a multi-domain knowledge base (Bizer et al., 2009). Until July 2011, the knowledge base had described more than 3.64 million things, of which 1.83 million were classified in a consistent ontology. In addition, it contains information about more than 60,000 companies\(^8\), so that we can link name entities by their categories, products, or services found in the knowledge source.

In order to recognize categories each name entity belongs to, we used the property: “dcterms:subject” to tag NEs with their categories. It’s because this property is used to describe the subject of a bibliographic resource\(^9\). For instance, Google has ten subjects in DBpedia data structure, including World_Wide_Web, Cloud_computing_providers, Web_service_providers, etc. Therefore, when two NEs are linked by concepts in a higher level of the semantic network, the relation between these two

---

\(^7\) From experiment, \( w_a \) and \( w_s \) are 0.5, respectively.

\(^8\) http://dbpedia.org

NEs are farther compared with two NEs linked by a more concrete concept, which usually lies in a lower level of the network (López-Nores et al., 2011). The ontological relation between NE, and NEj is calculated as $D_{ij} = \sum_{1 \leq l \leq n} (0.5d_l + 0.5d_j)$, where $D_{ij}$ represents distance between NE, and NEj in DBpedia semantic network. $d_i$ is the distance from the concept in the lowest level shared by NE, and NEj to NEi, and $d_j$ is the distance between the shared concept to NEj.

4 SYSTEM IMPLEMENTATION AND EXPERIMENTAL DESIGN

The main purpose of the research is to help users identify relations between business entities, technology, and product by filtering information with a personalization mechanism. Thus, there are three evaluation criteria for the proposed information retrieval mechanisms: (1) information quality of the search results; (2) the effectiveness of information retrieval in terms of discovering business relations; (3) the moderation effect contributed by the knowledge map.

We compared three information retrieval mechanisms on the aforementioned three criteria in order to evaluate the performance of the the proposed system. For evaluating information quality, we asked experts to identify news articles and relations in a specific topic. For evaluating the effectiveness of information retrieval, we adopted the task-based experimentation approach, asking subjects to answer questions as the task of business analysis in order to test if there is a significant difference in the number of business relations discovered by users using different information retrieval mechanisms. Moreover, by assessing the performance of question answering tasks, we tested the moderation effect of the knowledge map.

4.1 System implementation

In order to verify the performance of the proposed method, we implemented three information retrieval mechanisms. The system interface is shown in Figure 2, in which search results are displayed on the left box and business ecological diagram, composed of name entities from search results, on the right. Red nodes in the ecosystem represent product, and blue nodes are companies. The three retrieval mechanisms focusing on different aspects respectively are described as follows:

1. Retrieval by search term matching (TM): TM acts as the baseline mechanism; it retrieves news articles having the search term given by the user, and ranks the search results by their publication time, from the latest to the oldest. It mainly considers popularity; that is, it retrieved the most relevant articles according to the contents of news articles.

2. Retrieval by browsing history (BH): Since cosine similarity has been the most popular mechanism to calculate content similarity, we used cosine similarity to calculate the level of similarity between news articles and user’s profile. A user profile is represented as a vector consisting of NEs in the news repository browsed by each user, and the value is assigned 1 if the NE is read by the user, otherwise assigned 0. A news articles is also represented as a vector consisting of NEs, in which an NE is assigned value 1 if the NE is in the article, otherwise, assigned 0. The similarity score is calculated by $\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$, where $A$ stands for the term vector of a news article, and $B$ is the vector of a user profile. This method focuses on personal interest, owing to the comparison between news content and user profile.

3. Retrieval by ecological relation (ER): This is the proposed mechanism in this research. It considers popularity and personal preference simultaneously.

4.2 Experimental design

Since this research was designed to facilitate business ecological information discovery, we retrieved 1165 news articles from April 1st, 2012 to May 15th, 2012, and 2075 relations were extracted. The information sources include Business Week, Fast Company, Tech Crunch, Engadget, NYTimes,
Venture Beat, Wired, ZDnet, DGTimes and RedHerring. We then adopted task-based experimentation using questions answering as the task for business analysts. To identify the effects of knowledge maps on discovering ecological information, we developed two questionnaires by asking questions related to tablet and mobile devices respectively, and anticipated to generate two types of knowledge maps. Table 1 lists these two questionnaires. We adopted such measures as precision, recall and F-measure (F) to evaluate the performance of the information retrieval mechanisms. They are defined as

\[
\text{precision} = \frac{|\text{relevant documents} \cap |\text{retrieved documents}|}{|\text{retrieved documents}|},
\]

\[
\text{recall} = \frac{|\text{relevant documents} \cap |\text{retrieved documents}|}{|\text{total relevant documents}|}, \text{ and}
\]

\[
F = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.
\]

![Figure 2. System interface](image)

<table>
<thead>
<tr>
<th>Questionnaire A</th>
<th>Questionnaire B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q1.</strong> Please identify companies, products, or technologies related to “tablet” according to news articles retrieved by the experimental system.</td>
<td><strong>Q3.</strong> Mastercard just announced the smart wallet service, please identify which companies also provide this kind of service, including the name of the service and the technology used according to news articles retrieved by the experimental system.</td>
</tr>
<tr>
<td><strong>Q2.</strong> Please identify companies, products, or technologies related to “social network” according to news articles retrieved by the experimental system.</td>
<td><strong>Q3.</strong> Mastercard just announced the smart wallet service, please identify which companies also provide this kind of service, including the name of the service and the technology used according to news articles retrieved by the experimental system.</td>
</tr>
<tr>
<td><strong>Q3.</strong> Please identify companies, products, or technologies related to “e-book” according to news articles retrieved by the experimental system.</td>
<td><strong>Q4.</strong> There are more and more cloud services available these days. Please identify companies, products, or technologies related to “cloud storage” according to news articles retrieved by the experimental system.</td>
</tr>
<tr>
<td><strong>Q4.</strong> Please identify companies, products, or technologies related to “digital TV” according to news articles retrieved by the experimental system.</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Questionnaires used in the experiments

To test the information quality, we asked two master students majored in information systems and one majored in technology management to serve as the experts to evaluate the retrieved news articles. First, the three experts rated the relevance between topic and news articles retrieved according to the topic, from score 1, very irrelevant, to score 7, very relevant. We then selected news with score higher than 4, implying that they are related to the corresponding topics. After related news articles were
chosen, we finally asked the experts to select relations in the news that were also related to the specific topics as the set of correct answers. According to these answers, we can calculate each mechanism’s $F$-measure and the rate of correctness.

Meanwhile, we asked 36 master students majored in technology management to answer questions listed in Table 1. Half of students were randomly assigned to answer questionnaire A, and the rest were assigned to answer questionnaire B. We then asked them to act themselves as business analysts. We first used 10 questions to determine whether they have ICT background knowledge. In each group of people having or not having background knowledge, we assigned subjects to use one of the three information retrieval mechanisms, until each mechanism has been tested by 3 subjects. Therefore, each mechanism will be used by 12 subjects, and each questionnaire will be answered by 18 subjects.

### 5 EXPERIMENTAL RESULTS

This section analyzes the evaluation results in information quality, the effectiveness of the proposed information retrieval system, and the moderation effect of users’ background on system performance.

#### 5.1 Information quality

The relevant news articles and relations selected by three experts are listed in Table 2, and the number of relations retrieved by each mechanism is shown in Table 3. From the retrieved information, we calculated precision, recall and $F$-measure for each retrieval mechanism as shown in Table 4.

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Topic</th>
<th>Tablet</th>
<th>Social network</th>
<th>E-book</th>
<th>Digital TV</th>
<th>Smart wallet</th>
<th>Cloud storage</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM</td>
<td>Relations found</td>
<td>74</td>
<td>122</td>
<td>86</td>
<td>126</td>
<td>117</td>
<td>92</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>26</td>
<td>45</td>
<td>19</td>
<td>44</td>
<td>23</td>
<td>63</td>
<td>37</td>
</tr>
<tr>
<td>BH</td>
<td>Relations found</td>
<td>28</td>
<td>41</td>
<td>29</td>
<td>35</td>
<td>44</td>
<td>41</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>14</td>
<td>26</td>
<td>15</td>
<td>13</td>
<td>19</td>
<td>35</td>
<td>20</td>
</tr>
<tr>
<td>ER</td>
<td>Relations found</td>
<td>52</td>
<td>50</td>
<td>34</td>
<td>44</td>
<td>48</td>
<td>41</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>30</td>
<td>39</td>
<td>22</td>
<td>22</td>
<td>18</td>
<td>35</td>
<td>28</td>
</tr>
</tbody>
</table>

*Note. TM denotes term matching; BH denotes browsing history; ER denotes ecological relation.*

Table 2. Expert’s selection

Table 4 shows the performance of different retrieval mechanisms. Among three mechanisms, ER obtained the highest $F$-measure and precision, while TM gets the highest recall. This result can be explained by that TM retrieved too many news articles for one query, average 103 relations for each topic. BH does not perform well either in recall or in precision; the possible reason is that $cosine$
similarity with binary value vectors generates too small result sets for users to infer. However, ER retrieved relevant information based on the relationship between NEs, so that it could find more relations while eliminating irrelevant news articles comparing with TM, indicating the decrease in information overload.

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Topic</th>
<th>Tablet</th>
<th>Social network</th>
<th>E-book</th>
<th>Digital TV</th>
<th>Smart wallet</th>
<th>Cloud storage</th>
<th>Avg.</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM</td>
<td>Polled Recall</td>
<td>36.1%</td>
<td>100%</td>
<td>52.8%</td>
<td>89.8%</td>
<td>79.3%</td>
<td>100%</td>
<td>76.3%</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>35.1%</td>
<td>36.9%</td>
<td>22.1%</td>
<td>34.9%</td>
<td>19.7%</td>
<td>68.5%</td>
<td>36.2%</td>
<td></td>
</tr>
<tr>
<td>BH</td>
<td>Polled Recall</td>
<td>19.4%</td>
<td>57.8%</td>
<td>41.7%</td>
<td>26.5%</td>
<td>65.5%</td>
<td>55.6%</td>
<td>44.4%</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>50%</td>
<td>63.4%</td>
<td>51.7%</td>
<td>37.1%</td>
<td>43.2%</td>
<td>85.4%</td>
<td>55.1%</td>
<td></td>
</tr>
<tr>
<td>ER</td>
<td>Polled Recall</td>
<td>41.7%</td>
<td>86.7%</td>
<td>61.1%</td>
<td>44.9%</td>
<td>62.1%</td>
<td>55.6%</td>
<td>58.7%</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>57.7%</td>
<td>78%</td>
<td>64.7%</td>
<td>50%</td>
<td>37.5%</td>
<td>85.4%</td>
<td>62.2%</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Performance evaluation for each mechanism

5.2 Effectiveness of the proposed ER retrieval mechanism

The effectiveness of ecological relation discovery is shown in Table 5, in which you can find the scores obtained by human subjects. Score in the table is the rate of identifying the correct name entity in the expert’s selection. Each retrieval mechanism was tested by 12 human subjects to obtain the mean and standard deviation of the scores and time used.

We also adopted ANOVA and t-test to evaluate the significance of difference among these three mechanisms. We found that there is a significant difference of subjects’ performance using these three information retrieval mechanisms according to ANOVA result under 95% confidence level. According to the t-test results, subjects using ER significantly outperform the other two mechanisms, while no significant difference between term matching (TM) and browsing behavior (BH).

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Score</th>
<th>Time used (min.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term Matching (TM)</td>
<td>Mean</td>
<td>.1913</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Std.</td>
<td>.1431</td>
</tr>
<tr>
<td>Browsing History (BH)</td>
<td>Mean</td>
<td>.1685</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Std.</td>
<td>.0527</td>
</tr>
<tr>
<td>Ecological Relation (ER)</td>
<td>Mean</td>
<td>.3415</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Std.</td>
<td>.1695</td>
</tr>
</tbody>
</table>

Table 5. Subject’s score and time spent on each mechanism

5.3 Moderation effect of domain background knowledge

We found that background knowledge has a significant influence on human subjects in answering questions according to t-test under 95% confidence level. However, the personal knowledge map does
not affect the effectiveness of identifying relations no matter which retrieval mechanism is applied according to t-test results.

The original purpose of testing user’s background knowledge is to examine its effect on user’s performance by operating different information retrieval mechanisms. However, user’s knowledge map is only built on the experience of answering the first two questions. We can thus hardly discriminate users with background knowledge from those who without. Maybe the distinction between with and without domain knowledge will be clearer if we have more time and tasks assigned to users to develop their knowledge maps. Due to the limitation of this experimentation, we cannot link a person having specific domain knowledge with his/her knowledge map, and this contributes to the results that the moderation effect of background knowledge on user’s performance for different mechanisms is insignificant. Nevertheless, we found that background knowledge has more significant effect on TM and BH than on ER. This may suggest that solely relying on retrieved popular news or personal attention to recommend articles for users is insufficient. For ER mechanism, biased by personal attention with retrieved popular news articles, users can identify relevant relations easier.

In summary, comparing information quality among three information retrieval mechanisms, TM performed the highest recall, for that it retrieved articles containing the search term without filtering out any information, which is unsuitable for finding relations across different articles. BH obtained a small set of search results by solely adopting Boolean value to represent NEs without adequate query expansion. ER outperformed the other two mechanisms in F-measure by considering the relevance between search term and cadidate news articles both in dynamic and static relations between NEs. We also found that ER mechanism obtained a better result when information was scattered in various news articles. Therefore, it facilitates exploratory search the best among these three mechanisms.

6 CONCLUSION AND FUTURE WORK

The proposed information retrieval mechanism integrating ecological relation network and spreading activation method is able to identify more relations between name entities with fewer retrieved news articles, and in turn to decrease the information processing cost. The proposed mechanism balances personal attention and popularity of the topic with weights tuning for corresponding aspect, and outperforms the mechanism solely emphasizing personal pattern or popularity in measuring the information quality of search results. In addition, according to the experimental results, searching with ecological relation networks performed well when relations were scattered in various news articles, which indicates the ability of the proposed mechanism in facilitating exploratory search for business ecological information.

However, owing to the experimental limitations, it’s hard to establish users knowledge maps, which precisely depicts users’ attention and expertise. The inability of verifying the moderation effect of domain background knowledge on the effectiveness of search results shed lights on the future work; that is, the future research can hence focus on how to trace a user’s browsing behavior in order to develop a precise knowledge map, in terms of describing user’s attention. After the knowledge map is built, collaborative filtering mechanism can be considered to improve the recommendation performance. Besides user’s knowledge map, there are lots of other ways to balance personal preference with popularity of search targets. This might be examined in details in the future as well. Furthermore, other than DBpedia, lots of knowledge bases of a diversity of domains are available online that can be used to organize abstract knowledge.

References


