Fuzzy Information Retrieval Techniques for Help-Seeking System

Yueh-Min Huang[1,†] **Yen-Hung Kuo**[2,†] **Jim-Min Lin**[3,‡] **Yu-Lin Jeng**[4,†] **Juei-Nan Chen**[5,†]

[1]huang@mail.ncku.edu.tw, {[2]keh, [4]jeng, [5]nan}@easylearn.org, [3]jimmy@fcu.edu.tw

[†] Department of Engineering Science, National Cheng Kung University, Tainan City, Taiwan

[‡] Department of Information Engineering and Computer Science, Feng Chia University, Taichung City, Taiwan

Abstract

The forum system is useful for knowledge sharing and help-seeking. However, the existed forums are having the same problem such like nobody to answer the given questions. In order to attack the problem, this study proposes a human-expert oriented forum architecture to perform efficient knowledge sharing. The system uses fuzzy information retrieval techniques to discover important discussion knowledge and actively invites human-experts who may answer the question to participate the discussion.

Keywords: forum system, fuzzy information retrieval, help-seeking, human-expert oriented, knowledge sharing.

1. Introduction

Nowadays, the online forum becomes a useful tool for problem solving. In [1], it indicates the advantages of user forum to individual user. The most important benefit is users can receive the tailored answers from peers by formulating their problem within their own words without using properly keywords. Additionally, [2] concludes five categories of benefits when seeking information knowledge from other people. Opposed to the advantages, [1] argues the disadvantages of traditional user forums as follows.

- Suppose there is no participant interested in discussion issue, the proposed discussion might remain unanswered.
- Questions may remain in discussion group for a long time before it be answered by people.
- The help-seeker usually has no idea with recognizing the given advices which are conflicting recommendations.
- Apart from the discussions, the advices or solutions might formulate within poor readability.

Aim to the given four disadvantages, this paper proposes an intelligent human-expert oriented forum system to perform efficient knowledge sharing. To this end the proposed system includes both search engine and discussion forum. The help-seeker can specific problem with natural language to the search engine. The search engine would parse the user defined query by stemming [3] and n-gram [4] algorithms, and uses the parsed terms and fuzzy information retrieval technology [5] to discovery the relevant discussions for help-seeker. Nevertheless, suppose the searching results can not fit in with user's need then the user defined natural language query would automatically post to forum for further discussing. We call the posted unsolved problem as unsolved issue, and users can discuss unsolved issue on the forum system. Additionally, the forum would actively invite the users who may solve unsolved issue to participate the discussion. The forum system analyzes the discussion behavior of each user, and models their behaviors as an Expert-Term Correlation Matrix. In particularly, the matrix records the users' knowledge strengths patterns, and it is helpful for finding out the users who may provide tailored answer to help-seeker. Moreover, we call the system selected users are human experts. The users in the discussion environment play two roles, the first, help-seekers who ask questions, and the second, human experts who answer the unsolved issues. For each discussed issue, people can vote to the discussion or to the replies, suppose the voting result overcomes the predefined threshold, we say it is a solved issue. Whenever a problem is solved, the system would feedback the solved issue to its repository to be a static knowledge for further help-seeking. Following that, Fig. 1 shows the architecture and the communications of the introduced forum system.

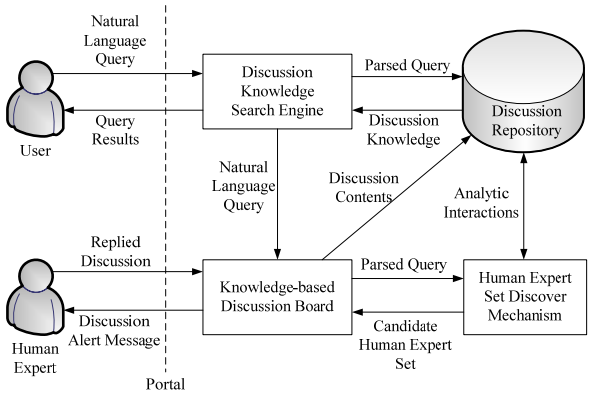

Fig. 1: The architecture of Human-Expert Oriented Forum System.

The rest of this article is organized as followings. Section 2 introduces our methodologies and its definitions. Section 3 draws our conclusions of this study.

2. Definitions and Methodologies

Section 2 gives some definitions and symbols to express the searching and matching mechanisms used in this study. Moreover, parts of the Fuzzy Information Retrieval techniques were expanded from [5]. In order to make this paper self-contained, we first illustrate the fundamental terms of Fuzzy Information Retrieval, and then express our methodologies.

2.1. Definitions

Definition 1: Discussion Set.

Let $D = \{d_1, d_2, \ldots, d_i\}$ be the set of discussions, where $\forall d_i \in D$, and $i > 0$.

Definition 2: Index Term Set.

Let k be the index term, then the discussion $d = \{k_1, k_2, \ldots, k_i\}$ can be presented by a set of index terms, where $i > 0$, and $\forall k_i \in d$. Moreover, an index term set is defined as $K = \cup(\text{members of } d_i)$, where $i > 0$, and $\forall d_i \in D$.

Definition 3: Query.

Let query $Q = \{q_1^{\pm}, q_2^{\pm}, ..., q_i^{\pm}\}$ be the set of query terms, where $\forall q_i^{\pm} \in Q$, and $i > 0$. In addition, for each query term $q_i^{\pm}$, it can be seen as a signed index term belongs to query discussion $d_q = \{k_1^{\pm}, k_2^{\pm}, ..., k_i^{\pm}\}$. Additionally, the q_i^{+} and q_i represent the positive and negative related query term respectively.

Definition 4: Human Expert Set.

Let $E = \{e_1, e_2, \ldots, e_i\}$ be the set of experts, where $\forall e_i \in E$, and $i > 0$.

Definition 5: TF*IDF.

Assume there are t index terms used to present a discussion d_j, and the d_j can be expressed as a vector $\vec{d}_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$, which $w_{i,j}$ is the ith weight of vector of d_j in t dimensions space. The $w_{i,j}$ can be seen as the importance of index term k_i to discussion d_j, and it can be estimated by formula (1). Referring to formula (1), it is similar to the famous tf*idf equation, the $freq_{i,j}$ is the frequency of index term k_i appears in discussion d_j. The $max_l freq_{l,j}$ is the number of most frequency index term k_l appears in discussion d_j. N is the number of discussions in D, and n_i is the number of discussions which contains index term k_i. In order to fit our applied, we add two parameters to the equation of term frequency (tf). For the two parameters, the details of its descriptions are shows as followings.

For the first parameter p, it presents the highest weight of term's occurrence positions in a discussion. The major idea is the different appearance of index terms should have different weights for presenting the positions.

For the second parameter r, it is a reply factor which indicates the additional importance of a term in a discussion. In the equation, count($reply_j$) presents the count of replies in a discussion d_j, and the count($reply_{i,j}$) presents the count of replies which contain the term k_i in discussion d_j. The idea for the parameter is that the important term should appear in each reply repeatedly.

$$w_{i,j} = tf_{i,j} * idf_i \qquad (1),$$

where $tf_{i,j} = p * \frac{freq_{i,j}}{max_l\, freq_{l,j}} * (1 + r)$,

$$idf_i = \log \frac{N}{n_i},$$

and $r = \begin{cases} \dfrac{\text{count}(reply_{i,j})}{\text{count}(reply_j)}, & \text{if has reply in } d_j \\ 0, & \text{if has no reply in } d_j \end{cases}$.

2.2. Terms Correlation Matrix

First of all, the thesaurus can be constructed based on $\vec{c}$ (Terms Correlation Matrix). The row and column is composed by the index terms in K. The coefficient $c_{i,l}$, which represents the interrelations between k_i and k_l, can be calculated by formula (2). For the formula (2), it uses conditional probability to exam the interrelationship between k_i and k_l. Referring to formula (2), the n_i stands for the number of discussions which include the index term k_i. Similarly,

$n_{i,l}$ means that the number of discussions which include both of k_i and k_l. Based on the $\vec{c}$, we can go a step further to build the degree of membership model from the discussion d_j to index term k_i.

$$c_{i,l} = \mathrm{P}(k_l|k_i) = \frac{n_{i,l}}{n_i} \qquad (2).$$

Making use of Definition 5, we can clean out an index term set *K*, and these index terms can represent the entire discussions in discussion set *D*. Therefore, we can collect these index terms to be the thesaurus about this topic, and construct the Fuzzy Term Expansion Model and the Expert-Term Correlation Matrix.

2.3. Fuzzy Term-Discussion Membership Degree

The degree of membership $\mu_{i,j}$ can be computed by using formula (3).

$$\mu_{i,j} = 1 - \prod_{k_l \in d_j} (1 - c_{i,l}) \qquad (3).$$

Let's examine an example to introduce the foundation idea of formula (3) and term expansion process. In some cases, the discussion d_j contains the index term k_l, but dose not contain the index term k_i. Therefore, we can ensure that $\mu_{l,j} \geq t_\mu$, where t_μ stands for the threshold of membership degree. In addition, $\mu_{l,j} \geq t_\mu$ means that index term k_l is highly related to discussion d_j, and we should put k_l to be an index term of d_j. The index terms k_i and k_l have the correlation coefficient $c_{i,l} \geq t_c$, where t_c stands for the threshold of the coefficient. Moreover, the $c_{i,l}$ satisfies t_c means that the index terms k_i and k_l have highly correlation. Suppose $\mu_{l,j} \geq t_\mu$ and $c_{i,l} \geq t_c$ cause the condition $\mu_{i,j} \geq t_\mu$ to become true, thus the index term k_i is related to the discussion d_j, even if the d_j does not contain k_i. Following that, the k_i should be added into the d_j's index term list. The entire process which expansions the d_j's index term list to include k_i is called the term expansion process.

2.4. Fuzzy Query Term Expansion

Based on formula (3), we can conduct formula (4) to present the query term expansion process. The query term expansion process can improve the recall rate due to it adds more information into query set.

$$\mu_{i,q} = 1 - \prod_{k_l \in d_q} (1 - c_{signed}),$$

$$\text{where } c_{signed} = \begin{cases} c_{i,l} & , \text{if } k_l = k_l^+ \\ (1 - c_{i,l}) & , \text{if } k_l = k_l^- \end{cases} \qquad (4)$$

Referring to formula (4), it assumes the set of query terms *Q* can be seen as a query discussion $d_q = \{k_1^\pm, k_2^\pm, ..., k_l^\pm\}$, and the terms involve in d_q are the members of *Q*. Given an index term $k_i \in K$, $k_i \notin d_q$, and it has coefficients $c_{i,p} \geq t_c$, $(1 - c_{i,n}) \geq t_c$ with query terms $k_p^+ \in d_q$, $k_n^- \in d_q$ respectively. Assume $c_{i,p} \geq t_c$ or $(1 - c_{i,n}) \geq t_c$ supports $\mu_{i,q} \geq t_\mu$ to become true, the index term k_i should be added into d_q to be the list of query terms. Finally, we denoted the expanded query set as the Q_{exp}.

2.5. Fuzzy Query

Given a query $Q = \{q_1^\pm, q_2^\pm, ..., q_i^\pm\}$, it consists of a set of positive and negative query terms. After query term expansion process, the query *Q* expands to Q_{exp}, and $Q \subseteq Q_{exp}$. Following that, it has to use Q_{exp} to find out the set of discussion relevant to *Q*. For the formula (5), the input discussion d_j has to estimate its score to the query Q_{exp} with multiply the membership degree between each query term k_q and discussion d_j. The estimated score of each discussion can be seen as a membership degree between discussion's index term set and expanded query term set. For the calculated scores, they can be used to rank the discussions, and the top of the ranked discussions would be the returned of query answer set.

$$score(d_j) = \prod_{k_q \in Q_{exp}} (\mu_{q,j} + \varepsilon),$$

$$\text{where } \mu_{q,j} = 1 - \prod_{k_l \in d_j} (1 - c_{signed}), \qquad (5)$$

$$c_{signed} = \begin{cases} c_{q,l} & , \text{if } k_q = k_q^+ \\ (1 - c_{q,l}) & , \text{if } k_q = k_q^- \end{cases}$$

2.6. Expert-Term Correlation Matrix

This section defines an Expert-Term Correlation Matrix for discovering a set of candidate human expert who can solve the given query question. Fig. 2 is a classic Expert-Term Correlation Matrix whose row is associated to a set of human experts $E = \{e_1, e_2, \ldots, e_j\}$, and the column is represented by a collected of index term set $K = \{k_1, k_2, \ldots, k_i\}$. In the matrix, it assumes each index term can be seen as a specific domain, and an expert may have a diversity of domain knowledge. Following that, each correlation between an index term k_i and an expert e_j is defined as strength $\delta_{i,j}$ ($\delta_{i,j} \geq 0$, and it is an integer), which represents the expert e_j has strength $\delta_{i,j}$ on the specific domain knowledge k_i (index term). Therefore, an expert e_j can

be expressed by a factor $\vec{e}_j = (\delta_{1,j}, \delta_{2,j}, ..., \delta_{i,j})$, where *i* is the number of index terms collected in *K*.

	e_1	e_2	…	e_j
k_1	$\delta_{1,1}$	$\delta_{1,2}$	…	$\delta_{1,j}$
k_2	$\delta_{2,1}$	$\delta_{2,2}$	…	$\delta_{2,j}$
⋮	⋮	⋮		⋮
k_i	$\delta_{i,1}$	$\delta_{i,2}$	…	$\delta_{i,j}$

Fig. 2: The Expert-Term Correlation Matrix.

In the Expert-Term Correlation Matrix, the strength $\delta_{i,j}$ is a dynamic variable which value is depending on the discussion behaviors of human expert e_j.

2.7. Candidate Human Expert Discovery

Fig. 3 shows the Human Expert Set Discover Mechanism, which combined with three main processes: 1) separating the positive expanded query term set (Q^+_{exp}) from expanded query (Q_{exp}), 2) according to the positive expanded query term set to calculate the strengths of each human expert to the query ($\Sigma\delta_{i,j}$), and 3) applying TOP_N() function to select top *N* strengths of experts to be the returned candidate human expert set (E_c). The details of the Human Expert Set Discover Mechanism would introduce as followings.

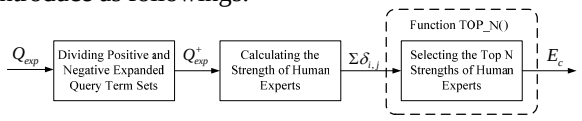

Fig. 3. Human Expert Set Discover Mechanism.

In order to discover the candidate expert set by using the Expert-Term Correlation Matrix, it has to separate all the positive query terms from an expanded query. The essential idea is it only cares about the expert who can solve the question, and the negative query set becomes unnecessary information. The Q^+_{exp} is the essential query information for generating candidate expert set E_c. formula (6) is used to discover E_c, the formula sums up each expert's strength which associated to Q^+_{exp} , and it selects highly strength of expert to be the candidate expert set.

$$E_c = \text{TOP\_N}(\Sigma\delta_{i,j}) \quad (6)$$

In formula (6), E_c is the candidate expert set, $E_c \subseteq E$. The $\delta_{i,j}$ is the correlation strength between k_i and e_j, where $k_i \in Q^+_{exp}$, and $e_j \in E$. Moreover, the TOP_N is the function, which returns the top *N* strength human experts.

3. Conclusions

This paper studies the entire knowledge delivering process on help-seeking system, and it gives the following two reasons for attacking the given disadvantages of traditional forum system. The first, the proposed forum system would actively invite the human experts to solve the given problem. Accordingly, questions would not fall in starvation circumstances, and the unsolved issue would become a solved one in a short time via human experts' interactions. Second, the information seekers can vote to discussions, therefore useful solutions' voting score would overcome the predefined threshold. These useful solutions passed peer validation and considered to become static knowledge in forum repository. Generally, help-seekers can trust these solved issues because they have higher confidences and passes peer review.

4. Acknowledgement

This work was supported in part by the National Science Council (NSC), Taiwan, R.O.C., under Grant NSC 94-2524-S-006-001.

5. References

[1] M.F. Steehouder, "Beyond Technical Documentation: User Helping Each Other," in *Proc. of IEEE International Professional Communication Conference*, Sept., pp. 489-499, 2002.

[2] R. Cross, R.E. Rice, and A. Parker, "Information seeking in social context: structural influences and receipt of information benefits," *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, Vol. 31, No. 4, pp. 438-448, 2001.

[3] M.F. Porter, "An algorithm for suffix stripping," *Program*, Vol. 14, No. 3, pp 130-137, 1980.

[4] P.F. Brown, V.J.D. Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer, "Class-Based n-gram Models of Natural Language," in *Proc. of the IBM Natural Language ITL*, pp. 283-298, 1990.

[5] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.