Topic Analysis for Psychiatric Document Retrieval

Liang-Chih Yu*‡, Chung-Hsien Wu*, Chin-Yew Lin†, Eduard Hovy‡ and Chia-Ling Lin*
*Department of CSIE, National Cheng Kung University, Tainan, Taiwan, R.O.C.
†Microsoft Research Asia, Beijing, China
‡Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA
liangchi@isi.edu,{chwu,totalwhite}@csie.ncku.edu.tw,cyl@microsoft.com,hovy@isi.edu

Abstract

Psychiatric document retrieval attempts to help people to efficiently and effectively locate the consultation documents relevant to their depressive problems. Individuals can understand how to alleviate their symptoms according to recommendations in the relevant documents. This work proposes the use of high-level topic information extracted from consultation documents to improve the precision of retrieval results. The topic information adopted herein includes *negative life events*, *depressive symptoms* and *semantic relations* between symptoms, which are beneficial for better understanding of users' queries. Experimental results show that the proposed approach achieves higher precision than the word-based retrieval models, namely the vector space model (VSM) and Okapi model, adopting word-level information alone.

1 Introduction

Individuals may suffer from negative or stressful life events, such as death of a family member, argument with a spouse and loss of a job. Such events play an important role in triggering depressive symptoms, such as depressed moods, suicide attempts and anxiety. Individuals under these circumstances can consult health professionals using message boards and other services. Health professionals respond with suggestions as soon as possible. However, the response time is generally several days, depending on both the processing time required by health professionals and the number of problems to be processed. Such a long response time is unacceptable, especially for patients suffering from psychiatric emergencies such as suicide attempts. A potential solution considers the problems that have been processed and the corresponding suggestions, called consultation documents, as the psychiatry web resources. These resources generally contain thousands of consultation documents (problem-response pairs), making them a useful information resource for mental health care and prevention. By referring to the relevant documents, individuals can become aware that they are not alone because many people have suffered from the same or similar problems. Additionally, they can understand how to alleviate their symptoms according to recommendations. However, browsing and searching all consultation documents to identify the relevant documents is time consuming and tends to become overwhelming. Individuals need to be able to retrieve the relevant consultation documents efficiently and effectively. Therefore, this work presents a novel mechanism to automatically retrieve the relevant consultation documents with respect to users' problems.

Traditional information retrieval systems represent queries and documents using a bag-of-words approach. Retrieval models, such as the *vector space model (VSM)* (Baeza-Yates and Ribeiro-Neto, 1999) and *Okapi model* (Robertson et al., 1995; Robertson et al., 1996; Okabe et al., 2005), are then adopted to estimate the relevance between queries and documents. The VSM represents each query and document as a vector of words, and adopts the *cosine measure* to estimate their relevance. The Okapi model, which has been used on the Text REtrieval Conference (TREC) collections, developed a family of word-weighting functions

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 1024–1031,
Prague, Czech Republic, June 2007.

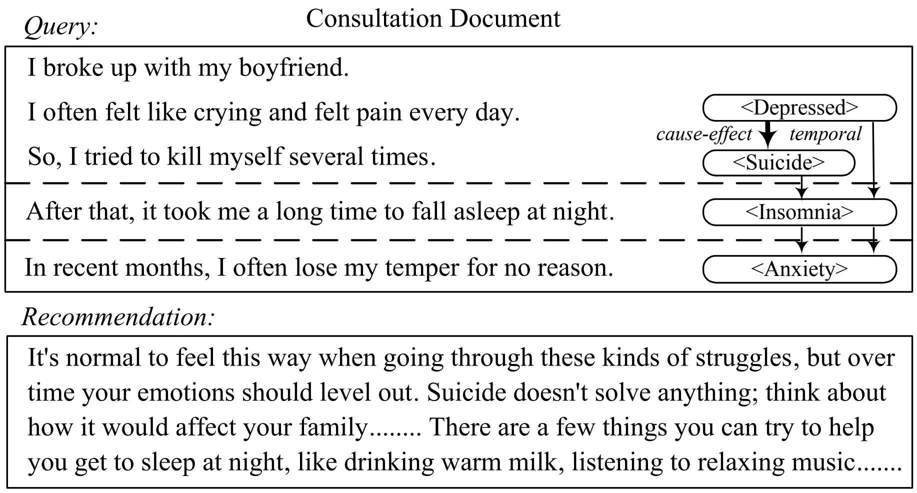

Figure 1. Example of a consultation document. The bold arrowed lines denote cause-effect relations; arrowed lines denote temporal relations; dashed lines denote temporal boundaries, and angle brackets denote depressive symptoms

for relevance estimation. These functions consider word frequencies and document lengths for word weighting. Both the VSM and Okapi models estimate the relevance by matching the words in a query with the words in a document. Additionally, query words can further be expanded by the concept hierarchy within general-purpose ontologies such as WordNet (Fellbaum, 1998), or automatically constructed ontologies (Yeh et al., 2004).

However, such word-based approaches only consider the word-level information in queries and documents, ignoring the high-level topic information that can help improve understanding of users' queries. Consider the example consultation document in Figure 1. A consultation document comprises two parts: the *query* part and *recommendation* part. The query part is a natural language text, containing rich *topic information* related to users' depressive problems. The topic information includes *negative life events*, *depressive symptoms*, and *semantic relations* between symptoms. As indicated in Figure 1, the subject suffered from a love-related event, and several depressive symptoms, such as <Depressed>, <Suicide>, <Insomnia> and <Anxiety>. Moreover, there is a cause-effect relation holding between <Depressed> and <Suicide>, and a temporal relation holding between <Depressed> and <Insomnia>. Different topics may lead to different suggestions decided by experts. Therefore, an ideal retrieval system for consultation documents should consider such topic information so as to improve the retrieval precision.

Natural language processing (NLP) techniques can be used to extract more precise information from natural language texts (Wu et al., 2005a; Wu et al., 2005b; Wu et al., 2006; Yu et al., 2007). This work adopts the methodology presented in (Wu et al. 2005a) to extract depressive symptoms and their relations, and adopts the pattern-based method presented in (Yu et al., 2007) to extract negative life events from both queries and consultation documents. This work also proposes a retrieval model to calculate the similarity between a query and a document by combining the similarities of the extracted topic information.

The rest of this work is organized as follows. Section 2 briefly describes the extraction of topic information. Section 3 presents the retrieval model. Section 4 summarizes the experimental results. Conclusions are finally drawn in Section 5.

2 Framework of Consultation Document Retrieval

Figure 2 shows the framework of consultation document retrieval. The retrieval process begins with receiving a user's query about his depressive problems in natural language. The example query is shown in Figure 1. The topic information is then extracted from the query, as shown in the center of Figure 2. The extracted topic information is repre-

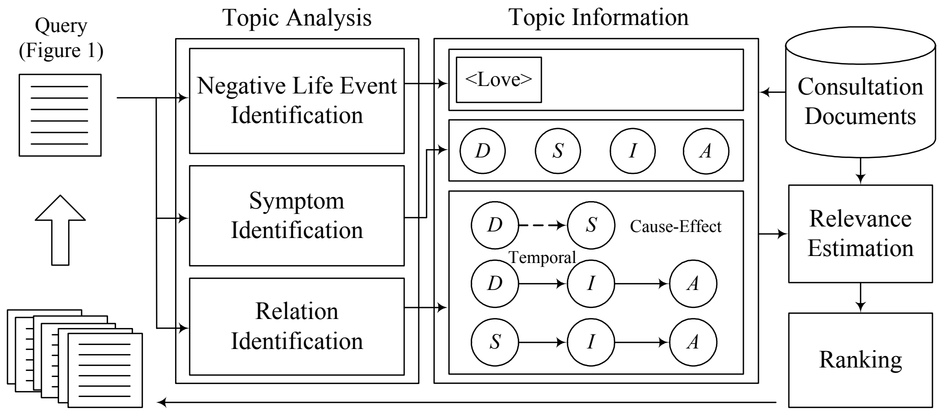

Figure 2. Framework of consultation document retrieval. The rectangle denotes a negative life event related to love relation. Each circle denotes a symptom. *D*: Depressed, *S*: Suicide, *I*: Insomnia, *A*: Anxiety.

sented by the sets of negative life events, depressive symptoms, and semantic relations. Each element in the event set and symptom set denotes an individual event and symptom, respectively, while each element in the relation set denotes a *symptom chain* to retain the order of symptoms. Similarly, the query parts of consultation documents are represented in the same manner. The relevance estimation then calculates the similarity between the input query and the query part of each consultation document by combining the similarities of the sets of events, symptoms, and relations within them. Finally, a list of consultation documents ranked in the descending order of similarities is returned to the user.

In the following, the extraction of topic information is described briefly. The detailed process is described in (Wu et al. 2005a) for symptom and relation identification, and in (Yu et al., 2007) for event identification.

1) **Symptom identification:** A total of 17 symptoms are defined based on the Hamilton Depression Rating Scale (HDRS) (Hamilton, 1960). The identification of symptoms is sentence-based. For each sentence, its structure is first analyzed by a probabilistic context free grammar (PCFG), built from the Sinica Treebank corpus developed by Academia Sinica, Taiwan (http://treebank.sinica.edu.tw), to generate a set of dependencies between word tokens. Each dependency has the format (*modifier*, *head*, $rel_{modifier,head}$). For instance, the dependency (matters, worry about, goal) means that "matters" is the goal to the head of the sentence "worry about". Each sentence can then be associated with a symptom based on the probabilities that dependencies occur in all symptoms, which are obtained from a set of training sentences.

2) **Relation Identification:** The semantic relations of interest include cause-effect and temporal relations. After the symptoms are obtained, the relations holding between symptoms (sentences) are identified by a set of discourse markers. For instance, the discourse markers "because" and "therefore" may signal cause-effect relations, and "before" and "after" may signal temporal relations.

3) **Negative life event identification:** A total of 5 types of events, namely <Family>, <Love>, <School>, <Work> and <Social> are defined based on Pagano et al's (2004) research. The identification of events is a pattern-based approach. A pattern denotes a semantically plausible combination of words, such as <parents, divorce> and <exam, fail>. First, a set of patterns is acquired from psychiatry web corpora by using an evolutionary inference algorithm. The event of each sentence is then identified by using an SVM classifier with the acquired patterns as features.

3 Retrieval Model

The similarity between a query and a document, $Sim(q,d)$, is calculated by combining the similarities of the sets of events, symptoms and relations within them, as shown in (1).

$$Sim(q,d) = \alpha Sim_{Evn}(q,d) + \beta Sim_{Sym}(q,d) + (1-\alpha-\beta) Sim_{Rel}(q,d), \tag{1}$$

where $Sim_{Evn}(q,d)$, $Sim_{Sym}(q,d)$ and $Sim_{Rel}(q,d)$, denote the similarities of the sets of events, symptoms and relations, respectively, between a query and a document, and α and β denote the combination factors.

3.1 Similarity of events and symptoms

The similarities of the sets of events and symptoms are calculated in the same method. The similarity of the event set (or symptom set) is calculated by comparing the events (or symptoms) in a query with those in a document. Additionally, only the events (or symptoms) with the same type are considered. The events (or symptoms) with different types are considered as irrelevant, i.e., no similarity. For instance, the event <Love> is considered as irrelevant to <Work>. The similarity of the event set is calculated by

$$Sim_{Evn}(q,d) = \frac{1}{N(Evn_q \cup Evn_d)} \sum_{e \in q \cap d} Type(e_q, e_d) \cos(e_q, e_d) + const., \tag{2}$$

where Evn_q and Evn_d denote the event set in a query and a document, respectively; e_q and e_d denote the events; $N(Evn_q \cup Evn_d)$ denotes the cardinality of the union of Evn_q and Evn_d as a normalization factor, and $Type(e_q, e_d)$ denotes an identity function to check whether two events have the same type, defined as

$$Type(e_q, e_d) = \begin{cases} 1 & Type(e_q) = Type(e_d) \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

The $\cos(e_q, e_d)$ denotes the cosine angle between two vectors of words representing e_q and e_d, as shown below.

$$\cos(e_q, e_d) = \frac{\sum_{i=1}^{T} w_{e_q}^{i} w_{e_d}^{i}}{\sqrt{\sum_{i=1}^{T} \left(w_{e_q}^{i}\right)^2} \sqrt{\sum_{i=1}^{T} \left(w_{e_d}^{i}\right)^2}}, \tag{4}$$

where w denotes a word in a vector, and T denotes the dimensionality of vectors. Accordingly, when two events have the same type, their similarity is given as $\cos(e_q, e_d)$ plus a constant, *const.*. Additionally, $\cos(e_q, e_d)$ and *const.* can be considered as the word-level and topic-level similarities, respectively. The optimal setting of *const.* is determined empirically.

3.2 Similarity of relations

When calculating the similarity of relations, only the relations with the same type are considered. That is, the cause-effect (or temporal) relations in a query are only compared with the cause-effect (or temporal) relations in a document. Therefore, the similarity of relation sets can be calculated as

$$Sim_{Rel}(q,d) = \frac{1}{Z} \sum_{r_q, r_d} Type(r_q, r_d) Sim(r_q, r_d), \tag{5}$$

$$Z = N_C(r_q) N_C(r_d) + N_T(r_q) N_T(r_d), \tag{6}$$

where r_q and r_d denote the relations in a query and a document, respectively; Z denotes the normalization factor for the number of relations; $Type(e_q, e_d)$ denotes an identity function similar to (3), and $N_C(\bullet)$ and $N_T(\bullet)$ denote the numbers of cause-effect and temporal relations.

Both cause-effect and temporal relations are represented by symptom chains. Hence, the similarity of relations is measured by the similarity of symptom chains. The main characteristic of a symptom chain is that it retains the cause-effect or temporal order of the symptoms within it. Therefore, the order of the symptoms must be considered when calculating the similarity of two symptom chains. Accordingly, a *sequence kernel function* (Lodhi et al., 2002; Cancedda et al., 2003) is adopted to calculate the similarity of two symptom chains. A sequence kernel compares two sequences of symbols (e.g., characters, words) based on the subsequences within them, but not individual symbols. Thereby, the order of the symptoms can be incorporated into the comparison process.

The sequence kernel calculates the similarity of two symptom chains by comparing their sub-symptom chains at different lengths. An increasing number of common sub-symptom chains indicates a greater similarity between two symptom chains. For instance, both the two symptom chains $s_1 s_2 s_3 s_4$ and $s_3 s_2 s_1$ contain the same symptoms s_1, s_2 and s_3, but in different orders. To calculate the similarity between these two symptom chains, the sequence kernel first calculates their similarities at length 2 and 3, and then averages the similarities at the two lengths. To calculate the similarity at length 2, the sequence kernel compares their sub-symptom chains of length 2, i.e., $\{s_1s_2, s_1s_3, s_1s_4, s_2s_3, s_2s_4, s_3s_4\}$ and $\{s_3s_2, s_3s_1, s_2s_1\}$. Similarly, their similarity at length 3 is calculated by comparing their sub-symptom chains of length 3, i.e., $\{s_1s_2s_3,\ s_1s_2s_4,\ s_1s_3s_4,\ s_2s_3s_4\}$ and $\{s_3s_2s_1\}$. Obviously, no similarity exists between $s_1s_2s_3s_4$ and $s_3s_2s_1$, since no sub-symptom chains are matched at both lengths. In this example, the sub-symptom chains of length 1, i.e., individual symptoms, do not have to be compared because they contain no information about the order of symptoms. Additionally, the sub-symptom chains of length 4 do not have to be compared, because the two symptom chains share no sub-symptom chains at this length. Hence, for any two symptom chains, the length of the sub-symptom chains to be compared ranges from two to the minimum length of the two symptom chains. The similarity of two symptom chains can be formally denoted as

$$
\begin{aligned}
Sim(r_q, r_d) &\equiv Sim(sc_q^{N_1}, sc_d^{N_2}) \\
&= K(sc_q^{N_1}, sc_d^{N_2}) \\
&= \frac{1}{N-1}\sum_{n=2}^{N} K_n(sc_q^{N_1}, sc_d^{N_2}),
\end{aligned}
\tag{7}
$$

where $sc_q^{N_1}$ and $sc_d^{N_2}$ denote the symptom chains corresponding to r_q and r_d, respectively; N_1 and N_2 denote the length of $sc_q^{N_1}$ and $sc_d^{N_2}$, respectively; $K(\bullet,\bullet)$ denotes the sequence kernel for calculating the similarity between two symptom chains; $K_n(\bullet,\bullet)$ denotes the sequence kernel for calculating the similarity between two symptom chains at length *n*, and *N* is the minimum length of the two symptom chains, i.e., $N = \min(N_1, N_2)$. The sequence kernel $K_n(sc_i^{N_1}, sc_j^{N_2})$ is defined as

$$
\begin{aligned}
K_n(sc_i^{N_1}, sc_j^{N_2}) &= \left\langle \frac{\Phi_n(sc_i^{N_1})}{\left\|\Phi_n(sc_i^{N_1})\right\|} \bullet \frac{\Phi_n(sc_j^{N_2})}{\left\|\Phi_n(sc_j^{N_2})\right\|} \right\rangle \\
&= \frac{\sum_{u \in SC^n} \phi_u(sc_i^{N_1})\phi_u(sc_j^{N_2})}{\sqrt{\sum_{u \in SC^n} \phi_u(sc_i^{N_1})\phi_u(sc_j^{N_1})}\sqrt{\sum_{u \in SC^n} \phi_u(sc_i^{N_2})\phi_u(sc_j^{N_2})}},
\end{aligned}
\tag{8}
$$

where $K_n(sc_i^{N_1}, sc_j^{N_2})$ is the normalized inner product of vectors $\Phi_n(sc_i^{N_1})$ and $\Phi_n(sc_j^{N_2})$; $\Phi_n(\bullet)$

Given the two symptom chains, $s_1s_2s_3$ and $s_1s_2s_3s_4$.
Their similarity is calculated as

$$K(s_1s_2s_3, s_1s_2s_3s_4) = \frac{1}{2}\sum_{n=2}^{3} K_n(s_1s_2s_3, s_1s_2s_3s_4).$$

$$
\begin{aligned}
K_2(s_1s_2s_3, s_1s_2s_3s_4) &= \left\langle \frac{\Phi_2(s_1s_2s_3)}{\left\|\Phi_2(s_1s_2s_3)\right\|} \bullet \frac{\Phi_2(s_1s_2s_3s_4)}{\left\|\Phi_2(s_1s_2s_3s_4)\right\|} \right\rangle \\
&= \frac{2+\lambda^2}{\sqrt{(3+2\lambda^2+\lambda^4)(2+\lambda^2)}} = 0.71(\lambda = 1)
\end{aligned}
$$

Similarly, $K_3(s_1s_2s_3, s_1s_2s_3s_4) = 0.5.$

	$\phi_{s_1s_2}$	$\phi_{s_1s_3}$	$\phi_{s_1s_4}$	$\phi_{s_2s_3}$	$\phi_{s_2s_4}$	$\phi_{s_3s_4}$	...	$\phi_{s_{16}s_{17}}$	
$\Phi_2(s_1s_2s_3) = ($	1	λ	0	1	0	0	...	0	)
$\Phi_2(s_1s_2s_3s_4) = ($	1	λ	λ^2	1	λ	1	...	0	)

$\Phi_2(s_1s_2s_3) \bullet \Phi_2(s_1s_2s_3s_4) = 2 + \lambda^2$

$\Phi_2(s_1s_2s_3) \bullet \Phi_2(s_1s_2s_3) = 2 + \lambda^2$

$\Phi_2(s_1s_2s_3s_4) \bullet \Phi_2(s_1s_2s_3s_4) = 3 + 2\lambda^2 + \lambda^4$

Figure 3. Illustrative example of relevance computation using the sequence kernel function.

denotes a mapping that transforms a given symptom chain into a vector of the sub-symptom chains of length *n*; $\phi_u(\bullet)$ denotes an element of the vector, representing the weight of a sub-symptom chain u, and SC^n denotes the set of all possible sub-symptom chains of length *n*. The weight of a sub-symptom chain, i.e., $\phi_u(\bullet)$, is defined as

$$
\phi_u(sc_i^{N_1}) = \begin{cases}
1 & u \text{ is a contiguous sub-symptom chain of } sc_i^{N_1} \\
\lambda^{\theta} & u \text{ is a non-contiguous sub-symptom chain} \\
 & \text{with } \theta \text{ skipped symptoms} \\
0 & u \text{ does not appear in } sc_i^{N_1},
\end{cases}
\tag{9}
$$

where $\lambda \in [0,1]$ denotes a *decay factor* that is adopted to penalize the non-contiguous sub-symptom chains occurred in a symptom chain based on the skipped symptoms. For instance, $\phi_{s_1s_2}(s_1s_2s_3) = \phi_{s_2s_3}(s_1s_2s_3) = 1$ since s_1s_2 and s_2s_3 are considered as contiguous in $s_1s_2s_3$, and $\phi_{s_1s_3}(s_1s_2s_3) = \lambda^1$ since s_1s_3 is a non-contiguous sub-symptom chain with one skipped symptom. The decay factor is adopted because a contiguous sub-symptom chain is preferable to a non-contiguous chain when comparing two symptom chains. The setting of the decay factor is domain dependent. If $\lambda = 1$, then no penalty is applied for skipping symptoms, and the cause-effect and temporal relations are transitive. The optimal setting of

Topic	Avg. Number
Negative Life Event	1.45
Depressive Symptom	4.40
Semantic Relation	3.35

Table 1. Characteristics of the test query set.

λ is determined empirically. Figure 3 presents an example to summarize the computation of the similarity between two symptom chains.

4 Experimental Results

4.1 Experiment setup

1) **Corpus:** The consultation documents were collected from the mental health website of the John Tung Foundation (http://www.jtf.org.tw) and the PsychPark (http://www.psychpark.org), a virtual psychiatric clinic, maintained by a group of volunteer professionals of Taiwan Association of Mental Health Informatics (Bai et al. 2001). Both of the web sites provide various kinds of free psychiatric services and update the consultation documents periodically. For privacy consideration, all personal information has been removed. A total of 3,650 consultation documents were collected for evaluating the retrieval model, of which 20 documents were randomly selected as the test query set, 100 documents were randomly selected as the tuning set to obtain the optimal parameter settings of involved retrieval models, and the remaining 3,530 documents were the reference set to be retrieved. Table 1 shows the average number of events, symptoms and relations in the test query set.

2) **Baselines:** The proposed method, denoted as Topic, was compared to two word-based retrieval models: the VSM and Okapi BM25 models. The VSM was implemented in terms of the standard TF-IDF weight. The Okapi BM25 model is defined as

$$\sum_{t \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} + k_2 \,|Q|\, \frac{avdl-dl}{avdl+dl}, \quad (10)$$

where t denotes a word in a query Q; qtf and tf denote the word frequencies occurring in a query and a document, respectively, and $w^{(1)}$ denotes the Robertson-Sparck Jones weight of t (without relevance feedback), defined as

$$w^{(1)} = \log \frac{N-n+0.5}{n+0.5}, \quad (11)$$

where N denotes the total number of documents, and n denotes the number of documents containing t. In (10), K is defined as

$$K = k_1((1-b) + b \cdot dl / avdl), \quad (12)$$

where dl and $avdl$ denote the length and average length of a document, respectively. The default values of k_1, k_2, k_3 and b are describe in (Robertson et al., 1996), where k_1 ranges from 1.0 to 2.0; k_2 is set to 0; k_3 is set to 8, and b ranges from 0.6 to 0.75. Additionally, BM25 can be considered as BM15 and BM11 when b is set to 1 and 0, respectively.

3) **Evaluation metric:** To evaluate the retrieval models, a multi-level relevance criterion was adopted. The relevance criterion was divided into four levels, as described below.
- Level 0: No topics are matched between a query and a document.
- Level 1: At least one topic is partially matched between a query and a document.
- Level 2: All of the three topics are partially matched between a query and a document.
- Level 3: All of the three topics are partially matched, and at least one topic is exactly matched between a query and a document.

To deal with the multi-level relevance, the *discounted cumulative gain (DCG)* (Jarvelin and Kekalainen, 2000) was adopted as the evaluation metric, defined as

$$DCG[i] = \begin{cases} G[1], & if \ i=1 \\ DCG[i-1] + G[i]/\log_c i, & \text{otherwise} \end{cases} \quad (13)$$

where i denotes the i-th document in the retrieved list; *G[i]* denotes the gain value, i.e., relevance levels, of the i-th document, and c denotes the parameter to penalize a retrieved document in a lower rank. That is, the DCG simultaneously considers the relevance levels, and the ranks in the retrieved list to measure the retrieval precision. For instance, let <3,2,3,0,0> denotes the retrieved list of five documents with their relevance levels. If no penalization is used, then the DCG values for

Relevance Level	Avg. Number
Level 1	18.50
Level 2	9.15
Level 3	2.20

Table 2. Average number of relevant documents for the test query set.

Retrieval Model	Avg. Time (seconds)
Topic	17.13
VSM	0.68
BM25	0.48

Table 4. Average query processing time of different retrieval models.

	DCG(5)	DCG(10)	DCG(20)	DCG(50)	DCG(100)
Topic	4.7516*	6.9298	7.6040*	8.3606*	9.3974*
BM25	4.4624	6.7023	7.1156	7.8129	8.6597
BM11	3.8877	4.9328	5.9589	6.9703	7.7057
VSM	2.3454	3.3195	4.4609	5.8179	6.6945
BM15	2.1362	2.6120	3.4487	4.5452	5.7020

Table 3. DCG values of different retrieval models. * Topic vs BM25 significantly different (p<0.05)

the retrieved list are <3,5,8,8,8>, and thus DCG[5]=8. Conversely, if c=2, then the documents retrieved at ranks lower than two are penalized. Hence, the DCG values for the retrieved list are <3,5,6.89,6.89,6.89>, and DCG[5]=6.89.

The relevance judgment was performed by three experienced physicians. First, the *pooling* method (Voorhees, 2000) was adopted to generate the candidate relevant documents for each test query by taking the top 50 ranked documents retrieved by each of the involved retrieval models, namely the VSM, BM25 and Topic. Two physicians then judged each candidate document based on the multilevel relevance criterion. Finally, the documents with disagreements between the two physicians were judged by the third physician. Table 2 shows the average number of relevant documents for the test query set.

4) **Optimal parameter setting:** The parameter settings of BM25 and Topic were evaluated using the tuning set. The optimal setting of BM25 were k_1 =1 and b=0.6. The other two parameters were set to the default values, i.e., $k_2 = 0$ and $k_3 = 8$. For the Topic model, the parameters required to be evaluated include the combination factors, α and β, described in (1); the constant *const.* described in (2), and the decay factor, λ, described in (9). The optimal settings were $\alpha = 0.3$; $\beta = 0.5$; *const.*=0.6 and $\lambda = 0.8$.

4.2 Retrieval results

The results are divided into two groups: the *precision* and *efficiency*. The retrieval precision was measured by DCG values. Additionally, a paired, two-tailed *t-test* was used to determine whether the performance difference was statistically significant. The retrieval efficiency was measure by the query processing time, i.e., the time for processing all the queries in the test query set.

Table 3 shows the comparative results of retrieval precision. The two variants of BM25, namely BM11 and BM15, are also considered in comparison. For the word-based retrieval models, both BM25 and BM11 outperformed the VSM, and BM15 performed worst. The Topic model achieved higher DCG values than both the BM-series models and VSM. The reasons are three-fold. First, a negative life event and a symptom can each be expressed by different words with the same or similar meaning. Therefore, the word-based models often failed to retrieve the relevant documents when different words were used in the input query. Second, a word may relate to different events and symptoms. For instance, the term "worry about" is a good indicator for both the symptoms <Anxiety> and <Hypochondriasis>. This may result in ambiguity for the word-based models. Third, the word-based models cannot capture semantic relations between symptoms. The Topic model incorporates not only the word-level information, but also more useful topic information about depressive problems, thus improving the retrieval results.

The query processing time was measured using a personal computer with Windows XP operating system, a 2.4GHz Pentium IV processor and 512MB RAM. Table 4 shows the results. The topic model required more processing time than both VSM and BM25, since identification of topics involves more detailed analysis, such as semantic parsing of sentences and symptom chain construction. This finding indicates that although the topic information can improve the retrieval precision, incorporating such high-precision features reduces the retrieval efficiency.

5 Conclusion

This work has presented the use of topic information for retrieving psychiatric consultation documents. The topic information can provide more precise information about users' depressive problems, thus improving the retrieval precision. The proposed framework can also be applied to different domains as long as the domain-specific topic information is identified. Future work will focus on more detailed experiments, including the contribution of each topic to retrieval precision, the effect of using different methods to combine topic information, and the evaluation on real users.

References

Baeza-Yates, R. and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley, Reading, MA.

Cancedda, N., E. Gaussier, C. Goutte, and J. M. Renders. 2003. Word-Sequence Kernels. *Journal of Machine Learning Research*, 3(6):1059-1082.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Hamilton, M. 1960. A Rating Scale for Depression. *Journal of Neurology, Neurosurgery and Psychiatry*, 23:56-62

Jarvelin, K. and J. Kekalainen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. *In Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41-48.

Lodhi, H., C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. 2002. Text Classification Using String Kernels. *Journal of Machine Learning Research*, 2(3):419-444.

Okabe, M., K. Umemura and S. Yamada. 2005. Query Expansion with the Minimum User Feedback by Transductive Learning. In *Proc. of HLT/EMNLP*, Vancouver, Canada, pages 963-970.

Pagano, M.E., A.E. Skodol, R.L. Stout, M.T. Shea, S. Yen, C.M. Grilo, C.A. Sanislow, D.S. Bender, T.H. McGlashan, M.C. Zanarini, and J.G. Gunderson. 2004. Stressful Life Events as Predictors of Functioning: Findings from the Collaborative Longitudinal Personality Disorders Study. *Acta Psychiatrica Scandinavica*, 110: 421-429.

Robertson, S. E., S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M.Gatford. 1995. Okapi at TREC-3. In *Proc. of the Third Text REtrieval Conference (TREC-3)*, NIST.

Robertson, S. E., S. Walker, M. M. Beaulieu, and M.Gatford. 1996. Okapi at TREC-4. In *Proc. of the fourth Text REtrieval Conference (TREC-4)*, NIST.

Voorhees, E. M. and D. K. Harman. 2000. Overview of the Sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1):3-35.

Wu, C. H., L. C. Yu, and F. L. Jang. 2005a. Using Semantic Dependencies to Mine Depressive Symptoms from Consultation Records. *IEEE Intelligent System*, 20(6):50-58.

Wu, C. H., J. F. Yeh, and M. J. Chen. 2005b. Domain-Specific FAQ Retrieval Using Independent Aspects. *ACM Trans. Asian Language Information Processing*, 4(1):1-17.

Wu, C. H., J. F. Yeh, and Y. S. Lai. 2006. Semantic Segment Extraction and Matching for Internet FAQ Retrieval. *IEEE Trans. Knowledge and Data Engineering*, 18(7):930-940.

Yeh, J. F., C. H. Wu, M. J. Chen, and L. C. Yu. 2004. Automated Alignment and Extraction of Bilingual Domain Ontology for Cross-Language Domain-Specific Applications. In *Proc. of the 20th COLING*, Geneva, Switzerland, pages 1140-1146.

Yu, L. C., C. H. Wu, Yeh, J. F., and F. L. Jang. 2007. HAL-based Evolutionary Inference for Pattern Induction from Psychiatry Web Resources. Accepted by *IEEE Trans. Evolutionary Computation.*