Bayesian Network based Sentence Retrieval Model

Keke Cai, Jiajun Bu*, Chun Chen, Kangmiao Liu, Wei Chen
College of Computer Science, Zhejiang University
Hangzhou, 310027, China
*Corresponding Author, +86 571 8795 1431
{caikeke, bjj, chenc, lkm, chenw}@zju.edu.cn

ABSTRACT

This paper makes an intensive investigation of the application of Bayesian network in sentence retrieval and introduces three Bayesian network based sentence retrieval models with or without consideration of term relationships. Term relationships in this paper are considered from two perspectives: relationships between pairs of terms and relationships between terms and term sets. Experiments have proven the efficiency of Bayesian network in the application of sentence retrieval. Particularly, retrieval result with consideration of the second kind of term relationship performs better in improving retrieval precision.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – Retrieval models.

General Terms:

Algorithms, Design, Performance, Experimentation

Keywords:

Sentence retrieval, Bayesian network, term relationship.

1. BAYESIAN NETWORK BASED SENTENCE RETRIEVAL MODELS

Sentence retrieval is to retrieve query-relevant sentences in response to users' queries. However, large amount of uncertainties involved in the process of sentence retrieval restrain the significant improvements in retrieval performance. In the field of information retrieval, Bayesian network [3] has been accepted as one of the most promising methodologies to deal with information uncertainty. Taking into account the intrinsic uncertainty of sentence retrieval, the advantage of incorporating Bayesian network into sentence retrieval is obvious.

Inspired by the idea above, a Bayesian network based sentence retrieval model (BNSR) is proposed. An example of the topology of BNSR retrieval model is shown in Figure 1. The relevance probability of sentence S_k to the query Q is evaluated as:

$$P(S_k \mid Q) = \sum_{T_i \in Pa(S_k)} w_{ik} * P(T_i \mid Q) \qquad (1)$$

$Pa(S_k)$ is defined as all terms in TS connecting to S_k, i.e., $Pa(S_k) = \{T_i \in TS \mid T_i \in S_k\}$; w_{ik} means the weight of term T_i in sentence S_k and is defined as: $w_{ik} = log(f_{S_k,T_i}) + 1$, here f_{S_k,T_i} represents the frequency of term T_i in sentence S_k; $P(T_i \mid Q) = 1$ if $T_i \in Q$ else $P(T_i \mid Q) = 1/M$, where M is the number of terms in the collection.

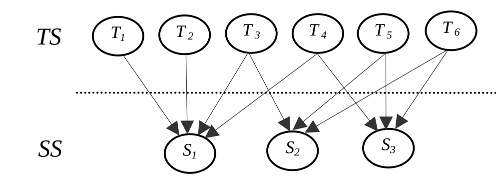

Figure 1. Topology of BNSR model.

In BNSR model, terms are assumed to be independent with each other. This assumption, although convenient in implementing, is not a reality. Term relationships deserve to be considered in the application of Bayesian network based sentence retrieval. Hence, this paper further proposes two expanded sentence retrieval models, i.e., BNSR_TR and BNSR_CR.

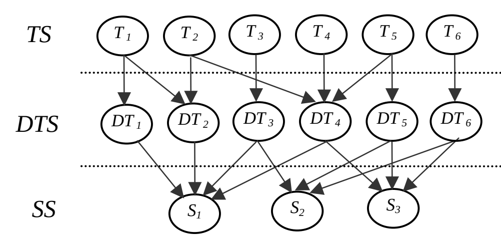

Figure 2. Topology of BNSR_TR model.

The main idea of BNSR_TR retrieval model is to utilize additional connections between different terms of query and sentence to facilitate the relevance identification of each sentence to query. An example topology of BNSR_TR model is shown in Figure 2. Compared with the BNSR model, BNSR_TR model has an additional term layer DTS that is constructed by duplicating each term in the term layer TS. Connections between terms of TS and DTS describe the relationships between pairs of terms. Here, the relationships are captured through an information space model, i.e., Hyperspace Analogue to Language (HAL) [2]. Given a term T_i in HAL, it can be represented by a n-dimensional term vector, each item describes the relevance of a term T_j to the term T_i and is formally described as $Rel_{T_i}(T_j)$. Based on this kind of term relationship, terms in DTS that are most relevant to each term in TS can be identified. Connections are then constructed by using arcs going from terms in TS to their relevant terms in DTS. Parents of term DT_j in DTS, or $Pa(DT_j)$, are terms of TS connecting to it.

Now, the relevance of sentence S_k to query Q can be evaluated through two steps: 1) compute the relevance probability of each term DT_j in DTS with respect to the query Q.

$$P(DT_j \mid Q) = \sum_{T_i \in Pa(DT_j)} u_{ij} * P(T_i \mid Q) \quad (2)$$

where u_{ij} equals to 1 if $DT_j = T_i$ otherwise $Rel_{T_i}(DT_j)$; 2) evaluate the relevance probability of S_k with respect to query Q.

$$P(S_k \mid Q) = \sum_{DT_j \in Pa(S_k)} w_{jk} * P(DT_j \mid Q) \quad (3)$$

Here, w_{jk} has the same definition as that in formula 1.

BNSR_TR incorporates term relationships into the inference process of retrieval, but ignores an important factor, the context, in which term relationships happen. Some inappropriate applications of term relationships are therefore incurred. To solve this problem, another expanded retrieval model BNSR_CR is proposed. An example of the topology of BNSR_CR retrieval model is shown in Figure 3. Compared with BNSR_TR, sentences in BNSR_CR are represented as a group of individual terms and terms sets. Term relationships are constructed between terms and term sets rather than between terms. Term set in this paper is defined as a frequent term set identified through frequency mining algorithm [1]. In general, the most advantage of this kind of relationship is that it reinforces the validities of those sentences identified relevant. In this paper, relevance between a term T_i and a term set TC_j, or $Rel_{T_i}(TC_j)$, is defined as the sum of association values between T_i and each term of TC_j. Based on this evaluation, term sets that are most relevant to terms in TS can be identified. Given a term T_i in TS, connections are then set up between T_i and each $TC_j \in TCS$, which either is relevant to T_i or equals to T_i .

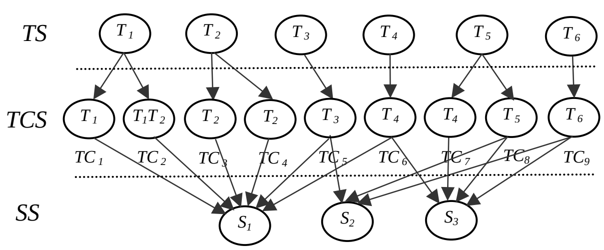

Figure 3. Topology of BNSR_CR model.

Now, the relevance probability of sentence S_k to query Q can be evaluated though the following computations: 1) evaluate the relevance probability of TC_j in TCS with respect to query Q:

$$P(TC_j \mid Q) = \sum_{T_i \in Pa(TC_j)} v_{ij} * P(T_i \mid Q) \quad (4)$$

where v_{ij} equals to 1 if $TC_j = T_i$ otherwise $Rel_{T_i}(TC_j)$; 2) evaluate the relevance probability of sentence S_k with respect to query Q:

$$P(S_k \mid Q) = \sum_{TC_j \in Pa(S_k)} w_{jk} * P(TC_j \mid Q) \quad (5)$$

Similarly, w_{jk} has the same definition as that in formula 1.

2. EXPERIMENTS

Our experiments are implemented on Aquaint Collection by using the TREC topics, N1-N100. Relevance of sentences that are retrieved is assessed by using the relevance assessments provided by TREC for the Novelty Task.

We compare the proposed retrieval models with three traditional approaches adopted for sentence retrieval: TFIDF model (TFIDF), OKAPI model (OKAPI) and KL-divergence model with Dirichlet smoothing (KLD). These three models are implemented by using the Lemur[1] toolkit. The comparison result in Table 1 and Table 2 show that the proposed sentence retrieval models outperform traditional retrieval models significantly. MAP represents the non-interpolated average precision averaged over all queries. AvgR is defined as C/R, where C is the number of the correctly identified sentences and R is the total number of relevant sentences for a given query, averaged over all queries. Moreover, the proposed retrieval models with consideration of term relationships perform better than that with no consideration of term relationships (BNSR). Experiment results of BNSR_TR and BNSR_CR also show that BNSR_TR performs better than BNSR_CR in improving retrieval recall while BNSR_CR performs better than BNSR_TR in improving retrieval precision.

Table 1. Performances of different models on topicsN1-N50

	TFIDF	OKAPI	KLD	BNSR	BNSR_ TR	BNSR_ CR
MAP	0.291	0.243	0.272	0.425	0.568	0.634
AvgR	0.607	0.575	0.592	0.643	0.886	0.798

Table 2. Performances of different models on topicsN51-N100

	TFIDF	OKAPI	KLD	BNSR	BNSR_ TR	BNSR_ CR
MAP	0.197	0.156	0.183	0.275	0.338	0.427
AvgR	0.639	0.605	0.626	0.681	0.878	0.804

3. CONCLUSIONS

This paper proposes three sentence retrieval models based on Bayesian network with or without consideration of term relationships. Experiments verify the performance improvements produced by Bayesian network based sentence retrieval approach. Particularly, the proposed retrieval models that take into consideration of term relationships perform better than that has no consideration of term relationships.

4. REFERENCES

[1] Grahne, G., Zhu, J. Efficiently using prefix-trees in mining frequent itemsets. *In proceedings of ICDM 2003 Workshop on Frequent Itemset Mining Implementations* (*FIMI'03*) (Melbourne, FL, USA, Dec. 19, 2003).

[2] Lund, K., Burgess, C. Producing High dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28, 2 (1996), 203-208.

[3] Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)

[1] http://www.lemurproject.org