A Survival Modeling Approach to Biomedical Search Result Diversification

Xiaoshi Yin[1,2], Jimmy Xiangji Huang [2], Xiaofeng Zhou [2], Zhoujun Li [1]
[1]School of Computer Science and Technology, Beihang University, Beijing, China.
[2]School of Information Technology, York University, Toronto, Canada.
xiaoshiyin@cse.buaa.edu.cn; jhuang@yorku.ca; lizj@buaa.edu.cn

ABSTRACT

In this paper, we propose a probabilistic survival model derived from the survival analysis theory for measuring aspect novelty. The retrieved documents' query-relevance and novelty are combined at the aspect level for re-ranking. Experiments conducted on the TREC 2006 and 2007 Genomics collections demonstrate the effectiveness of the proposed approach in promoting ranking diversity for biomedical information retrieval.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval
General Terms: Performance, Experimentation

Keywords: Survival Modeling, Diversity, Biomedical IR

1. INTRODUCTION

In the biomedical domain, the desired information of a question (query) asked by biologists usually is a list of a certain type of entities covering different aspects that are related to the question [6], such as genes, proteins, diseases, mutations, etc. Hence it is important for a biomedical information retrieval (IR) system to provide comprehensive and diverse answers to fulfill biologists' information needs. In the TREC 2006 and 2007 Genomics tracks, the "aspect retrieval" was investigated. Its purpose was to study how a biomedical IR system can support a user gather information about the different aspects of a topic. Aspects of a retrieved passage could be a list of named entities or MeSH terms, representing answers that cover different portions of a full answer to the query. Aspect Mean Average Precision (Aspect MAP) was defined in the Genomics tracks to capture similarities and differences among retrieved passages. It is a measurement for diversity of the IR ranked list [5].

The Genomics aspect retrieval was firstly proposed in the TREC 2006 Genomics track and further investigated in the 2007 Genomics track. However, to the best of our knowledge, there is not too much previous work conducted on the Genomics aspect retrieval for promoting diversity in the ranked list. University of Wisconsin re-ranked the retrieved passages using a clustering-based approach named GRASSHOPPER to promote ranking diversity [4]. Unfortunately, for the Genomics aspect retrieval, this re-ranking method hurt their system's performance and decreased the Aspect MAP of the original results [4]. Later in the TREC 2007 Genomics track, most teams tried to obtain the aspect level performance through their passage level results, instead of working on the aspect level retrieval directly [3, 6].

In this paper, we first propose a survival modeling approach to measuring the novel information provided by an aspect with respect to its occurrences. Then, the relevance and novelty of a retrieved document are combined at the aspect level. Evaluation results show that the proposed approach is effective in biomedical search result diversification.

2. SURVIVAL MODELING AND ANALYSIS FOR MEASURING NOVELTY

Survival analysis is a statistical methodology used for modeling and evaluating survival data, also called time-to-event data, where one is interested in the occurrence of events [2]. Survival time refers to a variable which measures the time from a particular starting time to a particular endpoint of interest. Events are usually referred as birth, death and failure that happen to an individual in the context of study. For example, in clinical trial, one may interested in the number of days that patient can survive in the study of the effectiveness of a new treatment for a disease. Formally, the survival function is defined as:

$$\begin{aligned} S(t) &= Pr(surviving\ longer\ than\ time\ t) \\ &= Pr(T > t) \end{aligned} \tag{1}$$

where t is a specific time, T is a random variable denoting the time of death, and "Pr" stands for probability. That is, the survival function gives the probability that the time of death is later than a specified time t. The survival function must be non-increasing: $S(u) \leq S(t)$ if $u > t$. Usually one assumes $S(0) = 1$, that is, at the start of the study, the probability of surviving past time zero is one. The survival function is also assumed to approach zero as t increases without bound [2].

In the context of information retrieval, aspects covered by a document can be considered as treatments, a document can be considered as a patient in the clinical trial case. The number of times that an aspect has been observed can be considered as the survival time. The new information that can be provided by an aspect corresponds to the effectiveness of a treatment. One can expect that, in a ranked list, as the number of times that an aspect has been observed increases, the new information that this aspect can provide to a document decreases. For example, in a ranked document list, when aspect "stroke treatment" is observed in the jth document at the first time, the information provided by this aspect should be counted as completely new. We presume that "stroke treatment" in the jth document covers the

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

topic of "medications taken by mouth for long-term stroke treatment". Then, when aspect "stroke treatment" is observed again in the kth ($k > j$) document of the ranked list, it may provide new information about "injection for short-term stroke treatment", but it is also possible that it only provides redundant information about "medications taken by mouth for long-term stroke treatment". As we can see, this situation satisfies the properties of the survival function described above.

We assume that the occurrences of an aspect follow Poisson distribution. Then, the survival model derived from Equation 1 can be formally written as:

$$S_{a_j}(x) = Pr(X > x) = 1 - e^{-\lambda} \sum_{i=0}^{x} \frac{\lambda^i}{i!} \qquad (2)$$

where λ is the rate parameter of Poisson distribution. The value of $S_{a_j}(x)$ states the probability of obtaining new information from aspect $a_j (j = 1, 2, ..., n$; where n is the number of observed aspects) after it has been observed x times. Note that in this paper, the aspects covered by a retrieved document are presented by concepts detected from Wikipedia [9].

3. COMBINING NOVELTY AND RELEVANCE

For retrieved documents, the document rankings should depend on which documents the user has already seen. Suppose that we have ranked top $i - 1$ documents, and now we need to decide which document should be ranked at the ith position in the ranking list. The document that can deliver the most new and relevant aspects should be considered as the ith document in the ranking list. Assume that aspect novelty and aspect query-relevance are independent of each other. Then given previous ranked $i - 1$ documents, we rank the ith document using the following scoring function:

$$\begin{aligned} score(d_i; d_1, ..., d_{i-1}) &= P(New\ and\ Rel|d_i) \\ &= \sum_{a_j \in A_{d_i}} P(New\ and\ Rel|a_j)P(a_j) \\ &\propto \sum_{a_j \in A_{d_i}} P(New|a_j)P(a_j|Rel) \qquad (3) \end{aligned}$$

where a_j is an aspect detected from document d_i, which follows Poisson distribution with an estimated rate parameter. $P(New\ and\ Rel|a_j)$ denotes the probability that a_j is query-relevant and can provide new information as well.

$P(New|a_j)$ in Equation 3 states the probability of obtaining new information from aspect a_j, which can be calculated using the survival models proposed in Section 2. $P(New|a_j) = 1$ when $i = 1$. Since we do not usually have relevance information, $P(a_j|Rel)$ is unavailable. One possible solution, as introduced in [7], is to consider that the best bet is to relate the probability of aspect a_j to the conditional probability of observing a_j given the query: $P(a_j|Rel) \approx P(a_j|Q)$. $P(a_j|Q)$ can be calculated by the two-stage model presented in [9].

4. EXPERIMENTAL RESULTS

We conduct a series of experiments to evaluate the effectiveness of the proposed model on the TREC 2006 and 2007 Genomics collections. For the 2007's topics, three baseline runs are used, which are NLMinter [3], MuMshFd [8] and an Okapi run [1]. NLMinter and MuMshFd were two of the most competitive IR runs submitted to the TREC 2007 Genomics track. For 2006's topics, we test our approach on three Okapi runs. In this paper, we mainly focus on the Aspect mean average precision (MAP), since our objective is to promote diversity in the IR ranked list.

Evaluation results of using the proposed approach for document re-ranking are shown in Table 1. The values in the parentheses are the relative rates of improvement over the original results. As we can see, our approach achieves promising and consistent performance improvements over all baseline runs. Performance improvements can be observed on both levels of evaluation measures. It is worth mentioning that our approach can further improve the best result (NLMinter) reported in the TREC 2007 Genomics track by achieving 18.9% improvement on Aspect MAP and 11% improvement on Passage2 MAP.

on 2007's topics			on 2006's topics		
MAP	Aspect	Passage2	MAP	Aspect	Passage2
NLMinter	0.2631	0.1148	Okapi06a	0.2176	0.0450
SvvModel	0.3117	0.1270	SvvModel	0.2379	0.0472
	(+18.5%)	(+10.6%)		(+9.3%)	(+4.9%)
MuMshFd	0.2068	0.0895	Okapi06b	0.3147	0.0968
SvvModel	0.2432	0.0926	SvvModel	0.3236	0.1009
	(+17.6%)	(+3.5%)		(+2.8%)	(+4.2%)
Okapi07	0.1428	0.0641	Okapi06c	0.2596	0.0601
SvvModel	0.1660	0.0669	SvvModel	0.2697	0.0624
	(+16.2%)	(+4.4%)		(+3.9%)	(+3.8%)

Table 1: Re-ranking Performance on Genomics topics

5. CONCLUSIONS

In this paper, we propose a survival modeling approach to promoting ranking diversity for biomedical information retrieval. The probabilistic survival model derived from the survival analysis theory measures the probability that novel information can be provided by an aspect with respect to its occurrences. Experimental results show that the proposed survival model can successfully capture the novel information delivered by aspects and achieve significant improvements on ranking diversity. We also show that combining the novelty and the relevance of a retrieved document at the aspect level is an effective way of promoting diversity of the ranked list, while keeping the relevance of retrieved documents. The proposed approach not only achieves promising performance improvements on the diversity based evaluation measure, but also on the relevance based evaluation measure.

Acknowledgements

This work is supported by NSERC Discovery Grant and an Early Researher Award of Ontario.

6. REFERENCES

[1] M. Beaulieu, M. Gatford, X. Huang, S. Robertson, S. Walker, and P. William. Okapi at TREC-5. In *Proc. of TREC-5*, 1997.

[2] D. Cox and D. Oakes. *Analysis of Survival Data.* Chapman Hall.

[3] D. Demner-Fushman and et. al. Combining resources to find answers to biomedical questions. In *Proc. of TREC-16*, 2007.

[4] A. B. Goldberg and et. al. Ranking biomedical passages for relevance and diversity: University of Wisconsin, Madison at TREC Genomics 2006. In *Proc. of TREC-15*, 2006.

[5] W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli. TREC 2006 Genomics track overview. In *Proc. of TREC-15*, 2006.

[6] W. Hersh, A. Cohen, L. Ruslen, and P. Roberts. TREC 2007 Genomics track overview. In *Proc. of TREC-16*, 2007.

[7] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceeding of the 24th ACM SIGIR.*

[8] N. Stokes and et. al. Entity-based relevance feedback for genomic list answer retrieval. In *Proc. of TREC-16*, 2007.

[9] X. Yin, X. Huang, and Z. Li. Promoting ranking diversity for biomedical information retrieval using wikipedia. In *Proc. of the 32nd ECIR*, 2010.