Parsimonious Concept Modeling

Edgar Meij[1] emeij@science.uva.nl
Dolf Trieschnigg[2] trieschn@ewi.utwente.nl
Maarten de Rijke[1] mdr@science.uva.nl
Wessel Kraaij[3] kraaijw@acm.org

[1]ISLA, University of Amsterdam The Netherlands
[2]HMI, University of Twente, Enschede, The Netherlands
[3]TNO ICT, Delft, The Netherlands

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Algorithms, Measurement, Theory, Experimentation

Keywords

Parsimonious Models, Language Models, Relevance Feedback

1. INTRODUCTION

In many collections, documents are annotated using concepts from a structured knowledge source such as an ontology or thesaurus. Examples include the news domain [7], where each news item is categorized according to the nature of the event that took place, and Wikipedia, with its per-article categories [1]. These categorizing systems originally stem from the cataloging systems used in libraries and conceptual search is commonly used in digital library environments at the front-end to support search and navigation. In this paper we want to employ the explicit knowledge used for annotation at the back-end, not just to improve retrieval performance, but also to generate high-quality term and concept suggestions. To do so, we use the dual document representation—concepts and terms—to create a generative language model for each concept, which bridges the gap between vocabulary terms and concepts. Related work has also used textual representations to represent concepts, see e.g., [1, 11], however, there are two important differences. First, we use statistical language modeling techniques to parametrize the concept models, by leveraging the dual representation of the documents. Second, we found that simple maximum likelihood estimation assigns too much probability mass to terms and concepts which may not be relevant to each document. Thus we apply an EM algorithm to "parsimonize" the document models.

The research questions we address are twofold: (i) what are the results of applying our model as compared to a query-likelihood baseline as well as compared to a run based on relevance models [9] and (ii) what is the influence of parsimonizing? To answer these questions, we use the TREC Genomics track test collections in conjunction with MedLine. MedLine contains over 16 million bibliographic records of publications from the life sciences domain and each abstract therein has been manually indexed by trained curators, who use concepts from the MeSH (Medical Subject Headings) thesaurus [10]. We show that our approach is able to achieve similar or better performance than relevance models, whilst at the same time providing high quality concepts to facilitate navigation. Examples will show that our parsimonious concept models generate terms that are more specific than those acquired through maximum likelihood estimates.

2. CONCEPTUAL QUERY MODELS

To integrate concepts in the retrieval process, we propose a conceptual query model which is an interpolation of the initial query with terms obtained from a double concept translation. In this translation, concepts are used as a pivot language [8]; the initial query is translated to concepts and back to expanded query terms:

$$\begin{aligned} P(t|Q) &= (1-\lambda)\cdot P_{\mathrm{ML}}(t|Q) + \lambda \cdot \textstyle\sum_{c\in\mathcal{C}} P(t|c,Q)P(c|Q) \\ &\approx (1-\lambda)\cdot \tfrac{\#(t,Q)}{|Q|} + \lambda\cdot \textstyle\sum_{c\in\mathcal{C}} P(t|c)P(c|Q), \end{aligned} \quad (1)$$

where $\#(t,Q)$ is the number of times term t occurs in query Q and $|Q|$ is the query length. Two components need to be estimated here: the probability of a concept given a query, $P(c|Q)$, and of a term given a concept, $P(t|c)$.

To acquire $P(t|c)$, we will use the assignments of MeSH concepts to documents in MedLine and aggregate over the documents $\mathcal{D}_c$ which are labeled with a particular concept c:

$$P(t|c) \quad = \quad \textstyle\sum_{D\in\mathcal{D}_c} P(t|D,c)P(D|c).$$

We drop the conditional dependence of t on c given a document D, assume $P(D)$ to be uniform, and apply Bayes' rule to obtain:

$$P(t|c) \quad = \quad \tfrac{1}{P(c)} \textstyle\sum_{D\in\mathcal{D}_\mathcal{C}} P(t|D)P(c|D), \quad (2)$$

where $P(c)$ is a maximum likelihood (ML) estimation on a background collection M:

$$P(c) = P(c|M) = \frac{\sum_D \#(c,D)}{\sum_{c'}\sum_{D'} \#(c',D')}.$$

However, if $P(t|D)$ and $P(c|D)$ are estimated based on ML, more general terms and concepts acquire too much probability mass, simply because they occur more frequently. To make the distributions more document specific, we consider both models to be a mixture of a document model $P(x|D)$ and a background model $P(x|M)$, where $x \in \{t, c\}$, and we "parsimonize" the ML estimate using the following EM algorithm [6]:

$$\text{E-step:} \qquad e_x = \#(x,D)\cdot \frac{\gamma P(x|D)}{(1-\gamma)P(x|M)+\gamma P(x|D)},$$

$$\text{M-step:} \qquad P(x|D) = \frac{e_x}{\sum_{x'} e_{x'}}. \quad (3)$$

For our experiments we fix $\gamma = 0.15$ [6]. Table 1a shows the effect of applying the parsimonious model to the estimation of concept D000544 ("Alzheimer Disease"). The parsimonious approach

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

a. “Alzheimer Disease”		b. Topic 186: “How do mutations in the Presenilin-1 gene affect Alzheimer’s disease?”	
MLE	Parsimonious	MLE	Parsimonious
disease	disease	Alzheimer Disease	**Presenilin-1**
alzheimers	**ad**	**Humans**	**Presenilin-2**
patients	alzheimers	Membrane Proteins	Alzheimer Disease
dementia	dementia	**Amyloid beta-Protein**	**Amyloid Precursor, Protein Secretases**
alzheimer	**amyloid**	Amyloid beta-Protein, Precursor	Membrane Proteins
brain	alzheimer	**Research Support, U.S. Gov’t, P.H.S.**	Amyloid beta-Protein, Precursor

Table 1: (a) Comparison of terms with the highest probability $P(t|c)$ for concept D000544: “Alzheimer Disease” and (b) a comparison of concepts with the highest probability $P(c|Q)$ for topic 186. Terms specific to a model are marked in boldface.

emphasizes more specific and thus more useful terms, including acronyms or abbreviations—“ad” in this particular example.

Next, we also need need a way of estimating concepts for each query, which means that we are looking for a set of concepts $\mathcal{C}_Q$ such that $c \in \mathcal{C}_Q$ have the highest posterior probability $P(c|Q)$. We approach this again by looking at the assignment of concepts to documents, but this time we consider documents which are related to the original query, by using the top ranked documents $\mathcal{D}_Q$ from an initial retrieval run:

$$P(c|Q) \;=\; \textstyle\sum_{D\in\mathcal{D}_Q} P(c|D)P(D|Q), \tag{4}$$

where $P(D|Q)$ is determined using the retrieval scores. Note that we assume that $P(c|D,Q) = P(c|D)$, such that we can directly use Eq. 3. A clear example of the effects of applying our model to the estimation of $P(c|Q)$ is given in Table 1b. The parsimonious approach is not only able to retrieve more specific concepts, such as “Presenilin-1”, but it is also able to retrieve multiple *aspects* of the topic, such as related genes, proteins, and diseases.

Test collection	Run	MAP	P@10	R-prec.
TREC Genomics 2004	QL	0.2799	0.4740	0.3138
	RM	**0.2976**	**0.5280**	**0.3307**
	CM	0.2911	0.4940	0.3251
TREC Genomics 2005	QL	0.2250	0.3898	0.2612
	RM	0.2274	0.3776	0.2595
	CM	**0.2338**	**0.3918**	**0.2639**
TREC Genomics 2006	QL	0.3562	0.4385	0.3625
	RM	0.3616	0.4462	0.3454
	CM	**0.3762**	**0.4538**	**0.3705**
TREC Genomics 2007	QL	0.2520	0.4000	0.2841
	RM	0.2487	0.3833	0.2687
	CM	**0.2582**	**0.4056**	**0.2877**

Table 2: Results for baseline query-likelihood run (QL), relevance models (RM), and conceptual query models (CM) (best results in boldface).

3. RESULTS AND DISCUSSION

To determine the retrieval performance of our conceptual query model, we compare it with a baseline query-likelihood run (QL) and a relevance feedback run based on Lavrenko and Croft [9]’s relevance models (RM) on the full range of available TREC Genomics test collections [2, 3, 4, 5]. We use the same document set $\mathcal{D}_Q$ ($|\mathcal{D}_Q| = 50$) and parameter settings for the RM runs and for our runs based on Eq. 1 (CM). The results of our experiments are listed in Table 2. (We did not perform extensive sweeps over possible values for $|\mathcal{D}_Q|$ or γ; we did explore λ and found that the optimal setting lies within the range 0.15–0.35.)

Although the differences in results are not statistically significant (between QL and RM, QL and CM, and RM and CM—tested using a two-tailed paired t-test at $p < 0.01$), we note that the conceptual query model and the relevance model consistently outperform the query-likelihood baseline. The only test collection where RM does not perform well is the 2007 collection, which may be the effect of the new task introduced that year [5]. CM thus rivals the performance of relevance models on most of the evaluated test collections, whilst it is able to generate sensible navigation suggestions in the form of relevant concepts.

4. CONCLUSION AND FUTURE WORK

We have introduced a parsimonious conceptual query model whose retrieval performance matches that of relevance models, while it is also able to generate high quality navigation suggestions in the form of concepts. Future work concerns further experimental validation of our results on additional test collections, as well as revisiting the modeling assumptions we made.

5. ACKNOWLEDGMENTS

This work was carried out in the context of the BioRange programme of the Netherlands Bioinformatics Centre, supported by a BSIK grant through the Netherlands Genomics Initiative, by the Virtual Laboratory for e-Science project, by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 017.001.190, 640.001.501, 640.002.501, STE-07-012 and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

6. REFERENCES

[1] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI’07*, 2007.

[2] W. Hersh, R. Bhuptiraju, L. Ross, P. Johnson, A. Cohen, and D. Kraemer. TREC 2004 Genomics track overview. In *TREC ’04*, 2004.

[3] W. Hersh, A. Cohen, J. Yang, R. Bhupatiraju, P. Roberts, and M. Hearst. TREC 2005 Genomics track overview. In *TREC ’05*, 2005.

[4] W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli. TREC 2006 Genomics track overview. In *TREC ’06*, 2006.

[5] W. Hersh, A. Cohen, and P. Roberts. TREC 2007 Genomics track overview. In *TREC ’07*, 2007.

[6] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *SIGIR ’04*, 2004.

[7] W.-L. Hsu and S.-D. Lang. Classification algorithms for netnews articles. In *CIKM ’99*, 1999.

[8] W. Kraaij and F. de Jong. Transitive probabilistic CLIR models. In *Proceedings of RIAO 2004*, 2004.

[9] V. Lavrenko and B. W. Croft. Relevance based language models. In *SIGIR ’01*, 2001.

[10] MedLine. `http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html#Medline`.

[11] D. R. Recupero. A new unsupervised method for document clustering by using WordNet lexical and conceptual relations. *Inf. Retr.*, 10(6): 563–579, 2007.