Comparison of methods for topic template queries in the biomedical domain

Antonio Jimeno-Yepes
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, UK
yepes@ebi.ac.uk

Rafael Berlanga-Llavori
Dept. Computer Systems and Languages
Universitat Jaume I
Castellon, Spain
berlanga@lsi.uji.es

Dietrich Rebholz-Schuhmann
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, UK
rebholz@ebi.ac.uk

Abstract

Topic template queries are focused on a facet of a structured user information need. Examples of these topic templates are: the role of gene G in disease D and the interaction of proteins P1 and P2. These templates allow for multiple instances and some commonalities might be found which might provide improved retrieval on unseen instance queries of a template.

In this paper, we have analyzed two possible solutions that integrate the analysis of existing results based on query reformulation and the boosting of documents based on text categorization.

We show that both approaches produce interesting results when enough example queries are provided and that the boosting of retrieved document based on text categorization has a better performance.

1 Introduction

In our work we are interested in topic template queries (TTQ). These queries are defined by a theme or subject which denotes a specific facet of related types of entities (e.g. the role of gene X in disease Y). Several instantiations of the templates are possible (e.g. the role of the *APC gene* in *colon cancer*). This setup might be useful in situations where there is a specific structured information need; e.g. researchers in the biomedical domain which are in charge of curating a database. In the next section we present the related work. Thereafter we introduce the methods we propose. Finally, we present the results and conclusions.

2 Related work

Our work is related to several approaches in IR that we briefly present in this section and compare to our problem.

Query reformulation might provide a better representation of the original user query. These techniques use the feedback obtained for each one of individual queries but do not optimize the search for unseen queries. Text classifiers build models for predefined categories (which represent a static information need). We can find techniques like scatter/gather where the documents retrieved by a search engine are organized into clusters, but which may not be relevant to the user.

Despite the variety of techniques for improving query results, there is no method that analyses the set of queries from a topic template to identify commonalities that might improve the retrieval performance of unseen queries given a topic template.

3 Methods

In this paper we compare two approaches that analyze explicit feedback in retrieval tasks for a set of queries and produce a model that improves the performance on unseen queries with the same topic template. The first approach is based on a boosting of retrieved documents according to a text categorizer that determines the relevance of the document to the topic template. The second one is a query reformulation approach.

3.1 Text categorization

This approach post-processes the result of an ad-hoc information retrieval system. A text classifier is applied on the top-n retrieved documents for a

Proceedings of the 3rd International Symposium on Languages in Biology and Medicine, Posters, pages 123-125, Jeju Island, South Korea, 8-10 November 2009.

given query and is used to boost documents that are deemed relevant. Since the text categorizers provide a different way of estimating relevance than the information retrieval system, the documents are boosted to the top of the retrieved list keeping the original rank among them. This is similar to the work of (Ruch et al., 2003) which combines a traditional vector space model and a rule based system. Several categorizers based on machine learning algorithms (Frank et al., 2005) with different learning bias have been compared: decision trees (J48), naïve bayes (NB), support vector machines (SMO) and k-nearest neighbors (K-NN). A cross-validation analysis is used to select the most adequate classifier for the task; given the algorithm and their possible parameters.

3.2 Query reformulation

The query reformulation used in the experiments is based on the query reformulation we proposed in (Jimeno-Yepes et al., 2009). We have modified our Ontology Query Model (OQM) (Jimeno-Yepes et al., 2009) to integrate the terms related the topic template. These terms are selected (learnt) from the relevance judgments provided in previous queries. We have introduced the topic template denoted by the relation $\mathcal{R}$ in a linear combination:

$$P(w_i|\mathcal{C},\mathcal{R}) = \alpha P_{CM}(w_i|\mathcal{C}) + \beta P_R(w_i|\mathcal{C}) + \gamma P_{Rel}(w_i|\mathcal{R}) \quad (1)$$

$$\alpha + \beta + \gamma = 1 \quad (2)$$

$P_{Rel}(w_i|\mathcal{R})$ depends only on the terminology linked to this relation and the terminology linked to other relations in the ontology. In the case of a richer relation ontology the probability would as well consider the occurrence of the terms in the other relations. Standard information retrieval statistics have been used to select the candidate terms.

4 Results

4.1 Experimental setup

The configuration of the system is the same we have used in (Jimeno-Yepes et al., 2009). The randomization test for paired data is used to compare statistically the methods († indicates $p < 0.01$). The training queries are used to retrieve the top-50 documents for each query. Documents are marked as positive or negative documents according to the benchmark. Random selection of negative documents is done to balance both classes. 5 times 2 fold cross validation is used to sample the set of queries for each data set due to the size of the data sets. Global results are the average of the results obtained for each one of the partitions.

4.2 Data sets

We have used two data sets for our experiments. One set considers the role of a gene in a disease and the second one the interaction of two proteins. These two data sets are presented in turn.

4.2.1 PGN-disease data set

We have used the 2005 TREC Genomics collection[1] because there is an interest on generic topics. This collection is made up of a subcollection of Medline, around 4M documents between years 1999 and 2004, and a collection of 50 queries. Queries are based on a topic template; i.e. the role of gene X in disease Y. From the TREC queries we have considered 20 queries related to the topic template.

4.2.2 PPI data set

We have used the DIP database [2], which deals with protein-protein interaction on yeast and has pointers to Medline articles. In total 260 queries are prepared. The average number of relevant documents per query is two. The document collection contains Medline citations till September 2004, about 15M Medline documents.

4.3 Identification of topical features results

4.3.1 PGN-disease data set results

The configuration for the PGN-disease data set is based on the results obtained from the relevance cleaning and refinement presented in (Jimeno-Yepes et al., 2009). The classifier with the best F-measure is SMO with an RBF kernel. We see a similar F-measure is obtained with a different trade off between precision and recall based on the capacity and the g parameter.

[1] http://ir.ohsu.edu/genomics/2005protocol.html

[2] http://dip.doe-mbi.ucla.edu/

As we can see in table 1, the best results are obtained with the baseline. This is because the training set does not allow finding a model that discards documents about the role of the PGN in the disease.

TREC	Rel. Retr	MAP
Baseline	747.2/1093.6	**0.3208**
Categorizer	747.2/1093.6	0.3149

Table 1: Refinement cleaning and categorization for PGN-disease

4.3.2 PPI data set results

The SMO classifier obtains the best F-measure result in the cross-validation analysis. The RBF kernel obtains a better performance. The method with highest recall is based on a linear kernel while the highest precision is obtained with the NB classifier.

From the list of terms identified for reformulation, there are terms that clearly denote an interaction like *interaction, binding, complex and hybrid*, terms that are related to experiments done to verify the interaction between proteins. These terms have been found relevant in a similar study by (Marcotte et al., 2001) and (Cohen et al., 2008). There are less obvious terms like *association* that have been found relevant in (Rebholz-Schuhmann et al., 2008).

In table 2, we present the result comparing the baseline methods with the modified ontology query model. The baseline methods are the co-occurrences based on cleaning and refinement (Jimeno-Yepes et al., 2009).

As we can see in table 2, both approaches perform better than the baseline. Boosting based on the text categorizer provides a better performance. This means that there are specific arrangements that the model produced by the SMO captures better than the query reformulation approach.

PPI	Rel. Retr	MAP
Baseline	189.2/317.2	0.1873
Categorizer	189.2/317.2	**0.2387**†
Refinement	**199.2/317.2**	0.2140 †

Table 2: Baseline, categorizer and refinement for PPI

5 Discussion

We have seen that the PPI set has the largest improvement over the baseline, compared to the PGN-disease set. If we analyze the query reformulation results, we see that in the PPI data set there is a common group of features that are repeated across the different folds and this explains the improvement over the baseline.

6 Future work

The categorization result indicates that there is a relation among the features that deserves further research. In addition, a normalization of the features using an ontology or a terminological resource as reference might reduce the sparcity of the feature set.

References

K.B. Cohen, M. Palmer, and L. Hunter. 2008. Nominalization and Alternations in Biomedical Language. *PLoS ONE*, 3(9).

E. Frank, M.A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I.H. Witten, and L. Trigg. 2005. Weka-a machine learning workbench for data mining. *The Data Mining and Knowledge Discovery Handbook*, pages 1305–1314.

A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann. 2009. Ontology refinement for improved information retrieval. *Information Processing & Management: Special Issue on Semantic Annotations in Information Retrieval (to appear)*.

E.M. Marcotte, I. Xenarios, and D. Eisenberg. 2001. Mining literature for protein-protein interactions. *Bioinformatics*, 17(4):359–363.

D. Rebholz-Schuhmann, A. Jimeno-Yepes, M. Arregui, and H. Kirsch. 2008. Assessment of Modifying versus Non-modifying Protein Interactions. In *The Third International Symposium on Semantic Mining in Biomedicine*.

P. Ruch, R. Baud, and A. Geissbhler. 2003. Learning-free text categorization. *9th Conference on Artificial Intelligence, in Medicine in Europe*.