Structure Cognizant Pseudo Relevance Feedback

Arjun Atreya V, Yogesh Kakde, Pushpak Bhattacharyya, Ganesh Ramakrishnan
CSE Department, IIT Bombay, Mumbai
{arjun,pb,ganesh}@cse.iitb.ac.in,yrkakde@gmail.com

Abstract

We propose a structure cognizant framework for pseudo relevance feedback (PRF). This has an application, for example, in selecting expansion terms for general search from subsets such as Wikipedia, wherein documents typically have a minimally fixed set of fields, *viz.*, *Title, Body, Infobox* and *Categories*. In existing approaches to PRF based expansion, weights of expansion terms do not depend on their field(s) of origin. This, we feel, is a weakness of current PRF approaches. We propose a per field EM formulation for finding the *importance* of the expansion terms, in line with traditional PRF. However, the final weight of an expansion term is found by weighting these *importance* based on whether the term belongs to the title, the body, the infobox or the category field(s). In our experiments with four languages, $viz.$, English, Spanish, Finnish and Hindi, we find that this structure-aware PRF yields a 2% to 30% improvement in performance (MAP) over the vanilla PRF. We conduct ablation tests to evaluate the importance of various fields. As expected, results from these tests emphasize the importance of fields in the order of title, body, categories and infobox.

1 Introduction

The ruling paradigm for Information retrieval (IR) (Manning et al., 2009) is *Pseudo Relevance feedback (PRF)*. In PRF, an assumption is made that the top retrieved documents are relevant to the query for picking expansion terms. Zhai and Lafferty (2001) show that using pseudo relevance feedback on monolingual retrieval improves the overall result considerably over the retrieval without PRF. In case of retrieval for languages with little web content, Chinnakotla et al., (2010) show that taking help of another language to expand query helps in better performance.

The motivation for our work is as follows. Every document in the web collection has certain structure associated with it $viz.$, title, body, links, $etc.$ Each of these fields has different level of importance in the document. For instance, document title broadly describes the whole document, whereas the body of the document contains the details. Content in these fields have different scales of contribution in uniquely representing that document in the collection. Hence it is important to consider the structure of a document while extracting expansion terms from it.

Structure based PRF, of course, draws on the basic theory of PRF as in Zhai and Lafferty (2001), which is based on expectation maximization (EM). We formulate a per field EM to get the weights of expansion terms and subsequently take their weighted sum in a spirit similar to mixture models.

2 Related Work

Approaches based on the use of external resources like wordnet for query expansion, though extensively studied, have been eventually dropped (Gong et al., 2005; Qiu and Frei, 1993). Several works have also used structure of documents for query expansion. These works propose the technique of first choosing relevant documents and finding expansion terms, therefrom, using cooccurence, meta tags *etc*. Al-Shboul and Myaeng (2011) use categories of Wikipedia pages to cluster documents and retrieve the relevant cluster for query. This approach gives better recall at the cost of precision.

Anchor texts in Wikipedia pages pointing to a category same as the query category are picked

International Joint Conference on Natural Language Processing, pages 982–986,
Nagoya, Japan, 14-18 October 2013.

as expansion terms in Ganesh and Verma (2009). This work exploits the structure only in the form of anchor texts and category information.

Techniques to disambiguate query terms based on disambiguation pages of Wikipedia are proposed in (Xu et al., 2009; Lin et al., 2010). Once disambiguated, the page is considered for picking expansion terms. Other literatures that deal with PRF based IR are (Milne. et al., 2007; Lin and Wu, 2008; Lv and Zhai, 2010; Jiang, 2011).

3 Our System

We make use of Wikipedia as an external document collection for picking expansion terms. Reasons for this are: *a*) open source *b*) well-defined structure *c*) authenticity due to crowdsourcing and review, *d*) coverage across domains and languages *e*) ever growing. Four fields from the Wikipedia document are considered $viz.$, *title, body, categories* and *infobox*.

Our problem statement is:

> *Given a query Q in a language L, retrieve relevant results from any document collection (WWW/dataset) in L using Wikipedia documents in L for generating expansion terms.*

The process of PRF based retrieval involves the following steps.

1. Retrieve ranked list of Wikipedia documents for a given query Q- $RetrievalModel$ (Section 3.1)
2. Pick expansion terms from the top k retrieved documents- $ExpansionModel$ (Section 3.2)
3. Obtain a modified query Q' by combining the expansion terms with the query terms- $AggregationModel$ (Section 3.3)
4. Retrieve ranked list of documents for the modified query Q'- $RetrievalModel$ (Section 3.1)

3.1 Retrieval Model

Language model based retrieval is used in (Ponte and Croft, 1998) and (Croft, 2003). For every document D, θ_D is the probability distribution of terms. Similarly, θ_Q is for the query Q. The "distance" between the query and a document, D_{KL} is calculated as equation 1.

$$D_{KL}\left(\theta_Q|\theta_D\right) = -\sum_{w} P(w|\theta_Q) log P(w|\theta_D) \quad (1)$$

The more the relevance of D, the less is $D_{KL}\left(\theta_Q|\theta_D\right)$.

3.2 Expansion Model

This model picks expansion terms that get combined with the query. Choosing expansion terms involves selecting a set of relevant documents and identifying terms that uniquely represent them. We use the retrieval model mentioned in section 3.1 to pick top k documents.

There exist many off-the-shelf expansion models to choose expansion terms from (Ganesh and Verma, 2009; Al-Shboul and Myaeng, 2011). None of these, however, exploit the structure of relevant documents. (Zhai and Lafferty, 2001) explain one of the state of art techniques to choose expansion terms using EM algorithm without considering the structure of a document. In Zhai and Lafferty (2001), a set of relevant documents R is retrieved and all terms in these documents are considered as observations. Since R is a subset of the document collection C, all terms in R also appear in C. Both R and C act as sources for generating terms.

Given a document, the content in each field of the document represents the document with different levels of importance. In our expansion model, we use Wikipedia as the source of expansion terms. Every Wikipedia document is composed of four fields *title, body, category* and *infobox*.

Expansion terms are picked independently from each field of the Wikipedia document. We run EM algorithm on each field as explained in Zhai and Lafferty (2001). We formulate an EM algorithm for pickng expansion terms from $Title$ field instead from a document as the whole. $Body$, $Categories$ and $Infobox$ fields follow the same formulation. The probability of all title terms in R ($P_{R_t k}$) is maximized using EM algorithm. Similarly, body terms, category terms and infobox terms are also maximized.

The output of interest in an iterative EM algorithm is the set of expansion terms for every field. EM algorithm gives the weights of the expansion terms, indicating their importance. Weighted combination of these sets of expansion terms from different fields of the document leads to the final set of expansion terms. Empirically decided weights (α's) are used for combining expansion terms from different fields as shown in the equa-

	Dataset	Query set	No.of documents
English	FIRE 2010	76-125(50)	1,25,586
Spanish	ELRA-E0036	41-200(160)	4,54,045
Finnish	ELRA-E0036	91-250(160)	55,344
Hindi	FIRE 2010	76-125(50)	95,216

Table 1: Details of Experimental Setup; numbers is parenthesis indicate the number of queries

	$NORF$	PRF	$StructPRF$
English	0.1758	0.2022 (+15%)	0.2189 (+24.5%)
Spanish	0.0433	0.1352 (+212%)	0.1778 (+310%)
Finnish	0.1532	0.2477 (+61.6%)	0.2517 (+64.3%)
Hindi	0.2321	0.2364 (+1.8%)	0.2529 (+9%)

Table 2: MAP scores; plus(+) indicates improvement over $NORF$

	$NORF$	PRF	$StructPRF$
English (2761)	1888	2080	2138
Spanish (2694)	391	1818	1919
Finnish (1377)	243	875	974
Hindi (915)	748	780	785

Table 3: Relevant documents retrieved; numbers in parenthesis indicate the actual relevant documents

tion 2. α_x indicates the importance given to the document field x.

$$P_{Rk} = \alpha_t \cdot P_{R_t k} + \alpha_b \cdot P_{R_b k} + \alpha_c \cdot P_{R_c k} + \alpha_i \cdot P_{R_i k} \quad (2)$$

where $\alpha_t + \alpha_b + \alpha_c + \alpha_i = 1$

3.3 Aggregation Model

Once expansion terms are picked from Wikipedia documents, they are merged with initial query terms. Introducing expansion terms increases the possibility of topic drift for the intended information need. Hence, it is important to give more weight to query terms compared to expansion terms. The equation 3 indicates the aggregation of query Q with the expansion terms E to get the modified query Q' with λ as the weight given to the query over the expansion terms.

$$Q' = \lambda Q + (1 - \lambda)E \quad (3)$$

4 Experimental Setup

We conduct experiments to evaluate the effect of document structure on expansion terms, using ELRA-E0036[1](part of CLEF) and FIRE 2010[2] datasets. Experiments are done in four languages, English, Spanish, Finnish and Hindi. Following are the set of experiments conducted:

NORF- No relevance feedback: This is the simplest form of retrieval without using any expansion.

PRF- Pseudo relevance feedback without using the structure of a document: This is traditional PRF. All terms in Wikipedia are considered to be equally important, and the naive expansion model of (Zhai and Lafferty, 2001) is used to find expansion terms.

StructPRF- Pseudo relevance feedback using the structure of a document: This is our proposed model. Structure of Wikipedia documents is used for finding expansion terms using the model described in section 3.2.

[1]http://catalog.elra.info/product_info.php?products_id=1127

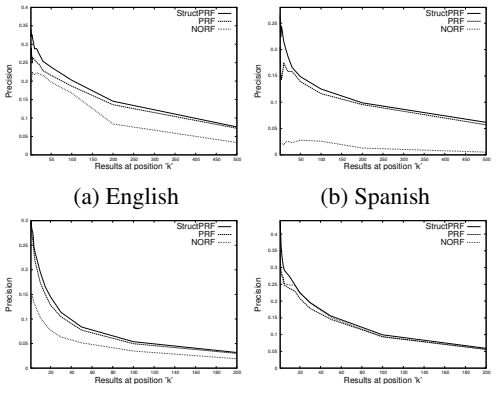
[2]http://www.isical.ac.in/∼fire/data.html

Table 1 describes the experimental details. For every query, 1000 results are retrieved and used for evaluation. All languages use their respective Wikipedia content for picking expansion terms.

5 Results

MAP scores are shown in table 2. $StructPRF$ has an overall improvement in MAP of 8% for English, 30% for Spanish, 2% for Finnish and 7% for Hindi over PRF. Figure 1 shows average precision values of all queries at different result positions for all languages. It is observed that there is a definite improvement in precision values for *StructPRF* over *PRF*. As we go down the list of retrievals ($P@k$, with k increasing), the improvement in *StructPRF* decreases but never gets below *PRF* and $NORF$.

Figure 2 depicts precision *vs.* recall curves for all languages. The results indicate that the *StructPRF* has a better precision for most recall val-



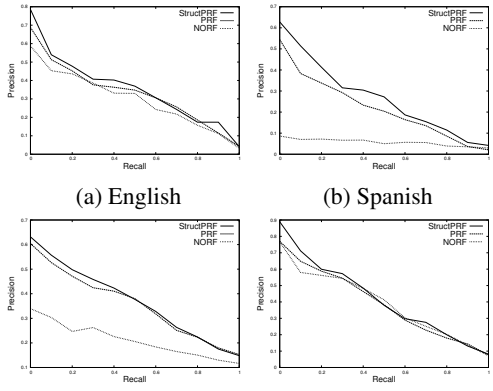
(a) English (b) Spanish

(c) Finnish (d) Hindi

Figure 1: $P@k$ Values

	English	Spanish	Finnish	Hindi
$NoTitle$	0.1953(-11%)	0.1179(-33%)	0.1914(-23%)	0.2086(-17%)
$NoBody$	0.2059(-6%)	0.1383(-22%)	0.2333(-8%)	0.2185(-13%)
$NoCategories$	0.2172(-0.7%)	0.1436(-19%)	0.2358(-7%)	0.2209(-12%)
$NoInfobox$	0.2178(-0.5%)	0.1467(-17%)	0.2449(-3%)	0.2234(-11%)

Table 4: MAP scores for ablation tests; minus(-) indicates percentage decrease from $StructPRF$



(a) English (b) Spanish

(c) Finnish (d) Hindi

Figure 2: Precision-Recall Curve

ues. At 60% to 80% recall, precision of *PRF* is better than *StructPRF* in English. This indicates that most of the relevant documents are pushed higher up the order in the result set. For Spanish and Finnish, $StructPRF$ consistently outperforms PRF. In Hindi, between 40% to 60% recall, PRF has a higher precision than $StructPRF$. This is again because of the relevant documents being pushed higher in the ranked list.

Analyzing query wise performances of $NORF$, PRF and $StructPRF$ for all languages, we observed that *StructPRF* has best precision compared to other two for ≈60% of queries in all languages.

Table 3 shows that there is an improvement in the number of relevant documents retrieved by *StructPRF* compared to PRF for all languages. $StructPRF$ has an improvement of 2.8%, 5%, 11% and 0.8% recall in English, Spanish, Finnish and Hindi respectively over PRF.

From these results it is evident that structure cognizant PRF benefits retrieval performance in terms of both precision and recall.

6 Ablation Tests

In ablation tests, we "disable" one field, that is, do not take expansion terms from a field, and get the MAP score. For instance, $NoTitle$ has body, categories and infobox with equal weights (*i.e.*, 1/3) and weight of the title field as 0.

Table 4 lists the MAP scores for all cases of ablation. The name of each of these cases indicates the field "disabled". It is observed that the worst degradation in MAP occurs on disabling the *Title* field. This happens for all languages. The degradation decreases in the order of *Title, Body, Categories* and *Infobox*.

The above observation translates to setting values for α_t, α_b, α_c and α_i described in section 3.2 as $\alpha_t > \alpha_b > \alpha_c > \alpha_i$ with $\alpha_t + \alpha_b + \alpha_c + \alpha_i = 1$. Hence the choice of α's for experimentation are 0.4, 0.3, 0.2 and 0.1 for α_t, α_b, α_c and α_i respectively.

The fields being important in the order of *Title, Body, Categories* and *Infobox* is quite intuitive. This is because the $Title$ represents the content of the document with a few words. Hence, the $Title$ field has a larger impact as compared to the $Body$ field. Though $Categories$ and $Infobox$ have lesser words, like $Title$, they refer to a generic context of the query.

7 Conclusions and Future Direction

In this paper, we have explored the usage of document structure for PRF. We proposed an expansion model that considers each field of the document with different levels of importance in picking expansion terms. This structure cognizant PRF is compared with both traditional PRF and with no-feedback, for four languages, English, Spanish, Finnish and Hindi. Experimental results show that using structure helps in getting considerable improvement in both precision and recall over traditional PRF. Ablation tests reveal the relative importance of the fields, with "title" field proving more important than others.

In our work, we combine expansion terms obtained from every field of a document in a decoupled way, that is, through separate per field EMs. In future, we would like to explore tight coupling of document fields (EM over individual per-field EM).

References

Bashar Al-Shboul and Sung-Hyon Myaeng. 2011. Query phrase expansion using wikipedia in patent class search. In *AIRS*, pages 115–126.

Manoj K. Chinnakotla, Karthik Raman, and Pushpak Bhattacharyya. 2010. Multilingual prf: english lends a helping hand. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 659–666, New York, NY, USA. ACM.

W Bruce Croft. 2003. Language models for information retrieval. In *Proceedings of 19th international conference on data engineering*, pages 3–7.

Surya Ganesh and Vasudeva Verma. 2009. Exploiting structure and content of wikipedia for query expansion in the context. In *International Conference RANLP*, pages 103–106.

Zhiguo Gong, Chan Wa Cheang, and U Leong Hou. 2005. Web query expansion by wordnet. In *In DEXA*, pages 166–175.

Xue Jiang. 2011. Query expansion based on a semantic graph model. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1315–1316, New York, NY, USA. ACM.

Tien-Chien Lin and Shih-Hung Wu. 2008. Query expansion via wikipedia link. In *ITIA'08:The 2008 International Conference on Information Technology and Industrial Application*.

Meng-Chun Lin, Ming-Xiang Li, Chih-Chuan Hsu, and Shih-Hung Wu. 2010. Query expansion from wikipedia and topic web crawler on clir. In *Proceedings of NTCIR-8 Workshop Meeting*, June 15-18.

Yuanhua Lv and ChengXiang Zhai. 2010. Positional relevance model for pseudo-relevance feedback. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 579–586, New York, NY, USA. ACM.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Inroduction to information Retrieval*. Cambridge University Press, Cambridge, England.

D Milne., Witten. I.H, and Nichols. D.M. 2007. A knowledge-based search engine powered by wikipedia. In *ACM Conference on Information and Knowledge Management*.

Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of 21st annual international ACM SIGIR conference on research and development in information retrieval*, pages 275–281.

Yonggang Qiu and Hans-Peter Frei. 1993. Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 160–169, New York, NY, USA. ACM.

Yang Xu, Gareth J.F. Jones, and Bin Wang. 2009. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 59–66, New York, NY, USA. ACM.

Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, New York, NY, USA. ACM.