Understanding Combination of Evidence using Generative Probabilistic Models for Information Retrieval

Paul Ogilvie
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
pto@lti.cs.cmu.edu

ABSTRACT

Structured documents, rich information needs, and detailed information about users are becoming more pervasive within everyday computing usage. Applications such as Question Answering, reading tutors, and XML retrieval demand more robust retrieval on richly annotated documents. In order to effectively serve these applications, the community will need a better understanding of the combination of evidence. In this work, I propose that the use of simple generative probabilistic models will be an effective framework for these problems. Statistical language models, which are a special case of generative probabilistic models, have been used extensively within recent Information Retrieval research. Their flexibility has been very effective in adapting to numerous tasks and problems. I propose to extend the statistical language modeling framework to handle rich information needs and documents with structural and linguistic annotations. Much of the prior work on combination of evidence has had few well-studied theoretical contributions, so I also propose to develop a more sound theoretical basis which gives more predictable results.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Theory

Keywords

language models, combination of evidence, structured documents, structured queries, user profiles

1. INTRODUCTION

As more complex information needs become more common, approaches that can leverage these rich information needs are becoming increasingly central for the future of Information Retrieval. An example of a context with rich information needs is a reading tutoring system, where reading difficulty and vocabulary to be stressed may be a part of the information need in addition to the standard topical component of information needs. These complex information needs present different sources of evidence about the properties of relevant information, and retrieval systems will need to effectively combine the different pieces of evidence in order to perform well in these environments.

Parallel to the increasing complexity of information needs, the availability of complex documents is increasing. Examples of complex documents include Internet web pages and XML documents. Both are rich in document structure and may have other sources of information beyond the standard text features used in traditional information retrieval (IR) systems. Another example of documents that have additional markup is documents used in a question answering (QA) system. These documents may have syntax and semantic markup. There is a need to develop models that can handle these rich information sources.

I'm in a unique position of having access to projects and data sets for each of the above examples. I will be actively working developing IR systems that can support these complex information needs in reading tutoring systems, XML retrieval, and serving documents ranked on richer queries to QA systems. The opportunity to work on real information needs for all of these problems is unique and will pose challenging problems to effective search.

A natural way to tackle the increasingly complex information needs and richer information sources is to treat these as different sources of evidence that need to be combined. Combination of evidence is widely recognized as an important problem within the field of Information Retrieval, but the conditions for success are not well understood. This thesis will strive to further understanding of these problems in the context of a retrieval method designed to handle numerous and diverse sources of evidence. I wish to model evidence from rich documents, rich information needs, and information about the user.

Understanding combination of evidence is becoming increasingly important as retrieval is becoming more varied in the forms of evidence and constraints. In this work, I will strive to understand the conditions for success when combining diverse sources of evidence. The framework in which I carry out this investigation is the generative probabilistic models for IR, which is a generalization of statistical language models. I will use this approach as I have found it an effective approach to combining evidence in prior research.

The generative proabilistic models are particularly amenable to combination of evidence, as the framework allows for natural combination of probability distributions estimated from different sources. Generalizing the statistical language modeling framework can allow the incorporation of text from various sources or even non-textual information. This will enable the retrieval system to not only consider a variety of document representations but additionally leverage different query representations and constraints within a single framework. Data analysis and the investigation of specific hypotheses will guide the development of appropriate modeling of the factors important for successful combination and aid in the development of smoothing methods that will enable effective combination techniques.

There are two main hypotheses of this work, each containing sub-hypotheses:

1. The statistical language modeling and more general generative probabilistic framework will allow for successful combination of evidence from document structure, query structure, and information about users.
    (a) Evidence from document structure can be modeled using generative distributions that can be combined into a mixture distribution representing the document.
    (b) Information about the users (for example, user interests) can be modeled as biases the rankings should exhibit. These biases will be expressed as posterior distributions desirable in the rankings.
    (c) Query constraints may be expressed as constraints on what the distributions generate, which source of evidence is used for generation, or as biases the rankings should exhibit. These constraints will also be expressed as posterior distributions desirable in the ranking. Some of these biases will be realized using prior probabilities, others through smoothing, parameter learning, or additional components of the query. The approach required will depend on the nature of the constraint, but there is not the space to describe this in detail here.
2. Successful combination of evidence will be dependent on the generative probability distributions for the combined models being compatible and appropriately weighted.
    (a) Compatible distributions are on the same scale. That is, a probability equal to x for a term in one distribution reflects the same degree of confidence of generation as a probability equal x for a term in another distribution.
    (b) Compatible distributions combine well. That is, when mixing two distributions, the resulting mixed distribution has probabilities that now accurately reflect the new degree of generation probability. This new distribution should be on the same scale as the distributions being combined.
    (c) The weighting of a distributions in a combination should reflect the quality or confidence of evidence in the distribution. The selection of appropriate weights may also depend on the amount of independent information present in a distribution.
    (d) The amount of independent quality information in the representations is a factor in how much benefit can be gained in the combination of evidence. There may be important connections to how the distributions vary for relevant and non-relevant documents.

The next section reviews related work. Section 3 describes previous work by the author in this area, and Section 4 describes how the author plans to investigate the hypotheses outlined above.

2. RELATED WORK

As mentioned earlier, there has been extensive work in the combination of evidence in the IR community. Much of this work is summarized by Croft [5]. Croft separates this work into several areas: combining representations, combining queries, combining ranking algorithms and search systems, and combining belief. In order to keep this discussion brief, I will only address work where a concerted effort was made to understand the conditions for successful retrieval.

As a result of early studies observing that different searchers forming queries for an information need created widely different queries, Belkin et al. [2] examined the conditions for successful combination of queries and found that bad query representations needed to be weighted lower than good query representations. They related this to the data fusion problem, which has been extensively studied for combining ranking algorithms and search systems.

Croft observed that combining ranking algorithms can be cast as problem of combining classifiers, which is well studied within the machine learning community. Tumer and Ghosh [16] provided a detailed analysis of using either linear combinations of classifiers or order statistics. Of particular interest to IR research is the linear combination of classifiers, as this is closely related to many of the combination techniques used within IR. Tumer and Ghosh showed that the combination of classifiers reduces the variance in boundary locations around the optimal boundary. They observed that classifiers of roughly similar quality combine the best, and including poorly performing classifiers can be detrimental to performance. Tumer and Ghosh also showed that the gain given by correlated classifiers is related to the amount of independent information expressed by the different classifiers.

An error in classification corresponds roughly to failing to retrieve a relevant document or retrieving a non-relevant document. However, combining rankings is not a simple matter of assigning a “retrieved” or “not-retrieved” label. The decision boundary varies based on thresholding of function or rank. Systems are evaluated with respect to multiple thresholds. How Tumer and Ghosh’s analysis can be generalized to the ranking problem is not entirely clear and has not been investigated. Additionally, Croft pointed out that Tumer and Ghosh’s work assumes the compatibility of the classifiers’ output, which is not always true for combining ranking algorithms.

Additional recent research has been done on the metasearch problem. Aslam and Montague [1] interpreted Croft's statements about combination of evidence by stating:

> "The systems being combined should (1) have compatible outputs, (2) each produce accurate estimates of relevance, and (3) be independent of each other."

However, these three hypotheses have not been fully investigated. A hypothesis similar to the independence hypothesis posed in metasearch research by Lee [7] is that there should be higher overlap of relevant documents (across the top n results given by each algorithm) than the overlap of non-relevant documents. However, Chowdhury et al. [3] found that this is not a sufficient condition for effective combination.

Manmatha et al. [8] took a different approach to metasearch. They explicitly modeled the distributions of relevant and non-relevant documents and used these as a guide for combination of rankings. While I do not expect that an explicit modeling of distributions will be necessary to do well in my own work, I do believe that modeling and understanding these distributions will be crucial understanding the conditions for successful combination of evidence.

In his discussion of combining belief, Croft described the INQUERY retrieval system which uses Bayesian inference networks. The inference network framework allows for multiple document and query representations (which may be structured queries). However, the framework gives little guidance on how to make sure the multiple forms of evidence can be combined successfully.

Greiff [6] described another probabilistic model that can incorporate multiple forms of evidence. Central to this work was the use of Exploratory Data Analysis (EDA) to model and understand factors important to relevance and successful combination of evidence. While I will not be working within the same probabilistic model as Greiff, I expect that EDA will be crucial within our own work to developing language models that can successfully combine evidence.

3. PREVIOUS WORK BY THE AUTHOR

I've done a variety of a variety of previous work[9][10][15] that has little relevance to this work, so I will not dwell on them here.

In the investigation of known item finding for the Web Track of the TREC Conference, I had the opportunity to do some research on the combination of evidence [12][14][4]. In this work, the different forms of evidence were give by document representations formed using the document structure of the HTML documents in the corpus. As a part of this research, I investigated some conditions for successful combination of information from the different document representations. This work investigated several hypotheses regarding combination of evidence from different search engines. A search engine was formed for each of the document representations. The problem of combining evidence was then that of the metasearch problem. I investigated metasearch hypotheses for combining evidence formed from document representations for known item retrieval. The investigation of these hypotheses resulted in several findings related to evidence combination, and I view this work as a useful preliminary investigation to factors that will be important for combining information.

In addition to the investigation of metasearch techniques, this work also examined a statistical language modeling approach to combining the document representations. The language modeling approach treated each document representation as a statistical language model. A new language model was formed for each document by taking a mixture of the document's language models. Documents were then ranked using the combined model. This approach was very successful at combining evidence from the different document representations, performing at least as well as metasearch approaches, and often outperformed the metasearch techniques. The implication of this on my future work is that language models may be an effective tool for combining information in other, more rich environments. There are still open questions in this work, such as why these distributions were compatible in this manner and how to weight the mixed models.

In [11], the author proposed a statistical language modeling framework for the retrieval of XML documents. The model proposed creates hierarchical language models from the structure of the XML documents. Smoothing of language models may be dependent on the node type or based on a general collection model. Document components for flat text queries are ranked by the probability that they generated the query. The model also describes ways to incorporate structured query constraints. Restrictions on the component type for query terms are modeled by limiting the generation of query terms to components of the appropriate type, where some query terms may be limited to children of a result. I view this work as some of the foundations for future work that I will continue to do with XML.

The author has also investigated a richer statistical language model for combining evidence from document structure present in XML documents [13]. This work was for adhoc retrieval of document components such as paragraphs or sections. The language model used here was the one presented above, but only for flat text queries. The simple interpolation methods used for this task at INEX performed quite well, demonstrating that richer methods for combining statistical language models can be effective.

4. PLAN

I've demonstrated that statistical language models can be an effective tool for combining evidence from document structure, and I will continue to adapt the framework to work for more complex information needs. Of particular interest is the problems I'll be working on. Some of these problems will be building back end retrieval engines that feed results into systems that can create complex queries and rich information needs.

The reading tutoring system mentioned in the Introduction (the REAP project) will provide a rich set of information needs. A typical information need may express the need for documents to be on a certain topic, within a certain range of reading difficulty levels, stress certain vocabulary the student is to learn, and be similar to other documents the student has viewed in the past. Future information needs in this environment may have additional constraints on the syntax structure of sentences in the document.

I will continue to work within the INEX community working on XML retrieval. The queries in this environment can be complex structured queries, and there is an abundance of relevance judgments on structured queries for structured document retrieval that will aid in the evaluation of my progress.

There will also be work on a QA project that will build more complex queries. Much of the work in QA has used simple queries in conjunction with basic retrieval systems, which return large result lists that for the QA systems to process. I will be developing a search engine that can handle more complex queries that have syntactic and semantic constraints on sentences in the documents. This will enable returning shorter result lists that have improved precision, but without sacrificing recall. An example query for "Who killed Abraham Lincoln?" would be something like return passages containing a sentence where the verb matches some language model representing the "kill" concept (may be generated from WordNet) and the object of the sentence matches the named-entity "Abraham Lincoln".

Some of the needs may be realized through prior probabilities, such as the desired reading level. Others may be treated as different query representations, such as matching a student's general interests (note that this could also be expressed as a prior probability). And others may be expressed as structured queries that constrain which language models are used for matching, as in a retrieval engine serving a QA system.

These very rich information needs will require successful combination of probability distributions, which stresses the importance of the hypotheses I listed in the Introduction. Some of the work I have done has begun to investigate some of these hypotheses, and I expect to continue work in this vein. That is, I plan to directly investigate the second set of hypotheses using exploratory and confirmatory data anaylsis. Much of the previous work in combination of evidence has recognized the importance of understanding how things work, but there have been few enlightening results. My preliminary experiments suggest that statistical language models tend to be more "well-behaved" than other models. As a result, I expect the generative probabilistic framework will be a valuable resource in the understanding of combination of evidence. In addition to gaining a better understanding of the combination of evidence, this work will also demonstrate how more complex information needs containing user profiles and constraints on the results can be incorporated into a retrieval system.

5. ACKNOWLEDGMENTS

This research was sponsored by National Science Foundation (NSF) grant no. CCR-0122581. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the NSF or the US government.

6. REFERENCES

[1] J. A. Aslam and M. Montague. Models for metasearch. In *Proc. of the 24th annual int. ACM SIGIR conf. on Research and development in information retrieval*, pages 276–284. ACM Press, 2001.

[2] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Inf. Process. and Management*, 31(3):431–448, 1995.

[3] A. Chowdhury, O. Frieder, D. Grossman, and C. McCabe. Analyses of multiple-evidence combinations for retrieval strategies. In *Proc. of the 24th annual int. ACM SIGIR conf. on Research and development in information retrieval*, pages 394–395. ACM Press, 2001.

[4] K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. P. Callan. Information filtering, novelty detection, and named-page finding. In *The Eleventh Text REtrieval Conf. (TREC-11), NIST SP 500-251*, pages 107–118, 2003.

[5] W. B. Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval*, pages 1–36. Kluwer, 2000.

[6] W. R. Greiff. The use of exploratory data analysis in information retrieval research. In *Advances in Information Retrieval*, pages 37–72. Kluwer, 2000.

[7] J. H. Lee. Analyses of multiple evidence combination. In *Proc. of the 20th annual int. ACM SIGIR conf. on Research and development in information retrieval*, pages 267–276. ACM Press, 1997.

[8] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *Proc. of the 24th annual int. ACM SIGIR conf. on Research and development in information retrieval*, pages 267–275. ACM Press, 2001.

[9] P. Ogilvie and J. P. Callan. The effectiveness of query expansion for distributed information retrieval. In *Proc. of the Tenth Int. Conf. on Information and Knowledge Management (CIKM-01)*, pages 183–190, New York, Nov. 5–10 2001. ACM Press.

[10] P. Ogilvie and J. P. Callan. Experiments using the lemur toolkit. In *The Tenth Text REtrieval Conf. (TREC-10), NIST SP 500-250*, pages 103–108, 2002.

[11] P. Ogilvie and J. P. Callan. Language models and structured document retrieval. In *Proc. of the First Annual Workshop of the Initiative for the Evaluation of XML retrieval (INEX)*, DELOS workshop, Dagstuhl, Germany, Dec. 2002. ERCIM.

[12] P. Ogilvie and J. P. Callan. Combining document representations for known-item search. In *Proc. of the 26th annual int. ACM SIGIR conf. on Research and development in informaion retrieval (SIGIR-03)*, pages 143–150, New York, July 28– Aug. –1 2003. ACM Press.

[13] P. Ogilvie and J. P. Callan. Using language models for flat text queries in xml retrieval. In *Proc. of the Second Annual Workshop of the Initiative for the Evaluation of XML retrieval (INEX) (to appear)*, Dagstuhl, Germany, Dec. 2003.

[14] P. Ogilvie and J. P. Callan. Combining structural information and the use of priors in mixed named-page and homepage finding. In *The Twelfth Text REtrieval Conf. (TREC-12) (to appear)*, 2004.

[15] L. Si, R. Jin, J. P. Callan, and P. Ogilvie. A language modeling framework for resource selection and results merging. In *Proc. of the Eleventh Int. Conf. on Information and Knowledge Management (CIKM-02)*, pages 391–397, New York, Nov. 4–9 2002. ACM Press.

[16] K. Tumer and J. Ghosh. Linear and order statistic combiners pattern classification. In A. Sharkley, editor, *Combining Artificial Neural Networks*, pages 127–162. Springer-Verlag, 1999.