A Unified Relevance Retrieval Model by Eliteness Hypothesis*

Jagadeesh Gorla[1], Stephen Robertson[2] and Jun Wang[1]
{j.gorla,j.wang}@cs.ucl.ac.uk, stephenerobertson@hotmail.co.uk

[1] University College London
[2] Microsoft Research, Cambridge

Abstract. In this paper, an Eliteness Hypothesis for information retrieval is proposed, where we define two generative processes to create information items and queries. By assuming the deterministic relationships between the eliteness of terms and relevance, we obtain a new theoretical retrieval framework. The resulting ranking function is a unified one as it is capable of using available relevance information on both the document and the query, which is otherwise unachievable by existing retrieval models. Our preliminary experiment on a simple ranking function has demonstrated the potential of the approach.

1 Introduction

In existing probabilistic retrieval models, the probability of relevance is calculated under two views. It is obtained either by correlating a given document with the information need properties of the users who would judge it relevant (conditioned by the given document), or by correlating the given information need with the information properties of those documents that they would be judged relevant (conditioned by the given query) [7]. The former is called *document-oriented view* and includes Maron and Kuhn's Probability Indexing and the language models [4, 5]; the latter is called *query-oriented view* and is represented by the Robertson-Spärck Jones model [8]. The two views rely on fixing one variable and optimising the other, i.e. conditioned by the information need and tuning the document or the other way around, but not both [6]. In this paper, we formulate a unified retrieval framework with the aim of using available relevance information on both the document and the query. Inspired by a 2-poisson model proposed in [2, 11], we state our Eliteness Hypothesis as:

"Any information, expressed by an author or a user, can be described using a binary set of elite properties. If we know the eliteness value of each elite property for a given information item (document) and need (query), we can deterministically determine whether the item-need pair is relevant or not."

We further assume that an author (or a user) will carry out the following process to express their information:

*A longer version can be found at http://arxiv.org/abs/1106.2946. See [3] also.

1. First, a user or author will choose a set of elite properties such that these properties can describe every aspect of the information that they want to express. The chosen elite properties are "elite" and the rest are "non-elite".
2. Once the elite properties are chosen, an observable information item or need, is generated by a stochastic function of chosen elite properties. The uncertainty about the eliteness is injected during the generation process.

Our hypothesis suggests that the relevance between document d and query q is known exactly if their elite properties, represented by $\mathbf{E}$, $\mathbf{F}$ respectively, are observed. Two deterministic relations are defined as follows:

A document d and query q are relevant if

- Strict identity: $\mathbf{E} = \mathbf{F}$, i.e. the eliteness of all the elite properties of d must be the same as the eliteness of the elite properties of q, or
- Logical inclusion: $\mathbf{F} \subset \mathbf{E}$, i.e. the elite properties of the document with eliteness "elite" must contain all the elite properties of the query whose eliteness is "elite".

The eliteness is, however, not observable and the relevance between d and q will be a probabilistic function given the generative process from Step 2. The unification is achieved by modifying the eliteness probabilities of the document-query pair when the relevance information is available on the document and/or the query, which is otherwise not provided by existing retrieval models.

2 The Probabilistic Relevance Ranking Function

We are ready to derive a probabilistic ranking function. Let D be a random variable whose value is any possible document generated by an author. Similarly, Let Q be a random variable whose possible values are any possible query generated by a user. We use the lower case d and q to denote their particular instantiations respectively. Let $\mathbf{E} \in \{0,1\}^k$ be a random binary vector over the space defined by the document's elite properties, where k is the number of elite properties, e.g. one elite property for each word in vocabulary. Similarly, $\mathbf{F} \in \{0,1\}^k$ is a random binary vector over the space defined by the query's elite properties. Their elements are denoted as E_i and F_i respectively, where $i \in \{1,k\}$. $E_i = 1$ means "elite",whereas $E_i = 0$ means "non-elite".

Our eliteness hypothesis gives the probability of relevance $P(R = 1|d,q)$ as

$$P(R = 1|d,q) = \sum_{\mathbf{E}} \sum_{\mathbf{F}} P(R = 1, \mathbf{E}, \mathbf{F}|d,q) \tag{1}$$

where R is a binary random variable. $R = 1$ means relevant while $R = 0$ otherwise. Applying Bayes' rule with independence assumptions[3] gives

$$P(R = 1|d,q) = P(R = 1) \sum_{\mathbf{E}} \sum_{\mathbf{F}} \prod_{i=1}^{k} \Big(\underbrace{\frac{P(E_i, F_i|R = 1)}{P(E_i)P(F_i)}}_{\text{part one}} \underbrace{P(E_i|d)P(F_i|q)}_{\text{part two}} \Big) \tag{2}$$

[3]Assumptions : 1) $\mathbf{E}$ is independent of q, $\mathbf{F}$ and similarly, $\mathbf{F}$ is independent of d, $\mathbf{E}$. 2) An elite property of a document or query is independent of other elite properties.

Eq. (2) uses the information about each elite property between the document and query in the relevant set ($P(E_i, F_i|R = 1)$). Thus, the information about other relevant document-query pairs that share the eliteness of elite property is included, which is an essential component of a unified model [7].

3 A Simple Ranking Function and its Experiments

In order to test the theory, we derive a very basic ranking function as an initial study. We assume the query terms to represent elite properties of the query ($F_i = 1$ when $q_i = 1$). Employing a simplified version of the logical inclusion hypothesis and ignoring equation terms that are ranking independent and terms with non-elite properties of the query result in the following ranking function:

$$P(R = 1|d, q) \propto \sum_{\forall i:F_i=1} \log \frac{P(E_i = 1|d)}{P(E_i = 1)} = \sum_{\forall i:F_i=1} \log \frac{P(tf_i|E_i = 1)}{\sum_{E_i \in \{0,1\}} P(tf_i|E_i)P(E_i)} \tag{3}$$

where $P(E_i = 1|d) = P(E_i = 1|tf_i)$ and tf_i denotes the term frequency of term i in document d. We refer to [3] for details. To compute Eq. (3), we adopt the 2-Poisson model in [2, 10]. That is the term frequency follows a Poisson distribution in the elite set of documents ($P(tf_i|E_i = 1) := e^{-\mu_i(1)}\mu_i(1)^{tf_i}$), and another Poisson distribution in the non-elite set ($P(tf_i|E_i = 0) := e^{-\mu_i(0)}\mu_i(0)^{tf_i}$), where $\mu_i(1)$ and $\mu_i(0)$ are the two Poisson means. The mixing probability $P(E_i = 1) := p_i$ is an additional parameter. Different with [2, 10], we obtained the optimal maximum likelihood parameters using the Expectation Maximization algorithm [1]. The detailed formulation of the EM algorithm can be found in [3]. By substituting the estimated parameter values, we get the final ranking function

$$P(R = 1|d, q) \propto \sum_{\forall i:F_i=1} log\Big(\frac{e^{-\mu_i(1)}\mu_i(1)^{tf_i}}{p_i e^{-\mu_i(1)}\mu_i(1)^{tf_i} + (1 - p_i)e^{-\mu_i(0)}\mu_i(0)^{tf_i}}\Big) \tag{4}$$

where the 2-Poisson model assumption assumes a fixed document length for all the documents [2, 11]. To circumvent this problem, we modify tf_i as $tf_i := tf_i\Big(b+(1-b)\frac{avgDL}{DL}\Big)$ where $avgDL$ is average document length in the collection, DL is the document length and $b \in [0, 1]$.

For the experiment, we employed the TREC-8 ad hoc task collection and topics to evaluate the ranking function in Eq. (4) against Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Recall@1000 metrics. We initialised the parameters with the data collection statistics. The initial value of p_i was set to be the portion of documents collection with the term-description property, $\mu_i(1)$ was initialised with the average number of times the term appeared in document with its term frequency more than one. We used a minuscule value to initialise $\mu_i(0)$. The assumption is that the average term frequency of a term associated with the term-description elite property in a document approaches zero if it is "non-elite" to the document. We have run tests and compared the performance of our method with BM25 [9] and Language Models [5] (with the

Model	MAP	MRR	Recall@1000
BM25	0.250	0.638	0.6634
Language Model with JM smoothing	0.238	0.4816	0.658
Language Model with Dirichlet prior	0.2539	0.6376	0.6694
Unified Model	0.2553 (0.2266*)	0.607 (0.6513*)	0.6659

Table 1: Performance on the TREC-8 ad hoc task data collection.

Jensen-Mercer smoothing and Dirichlet prior). Table 1 shows that our method outperforms the baselines when directly optimising the metrics. The improvement over MRR is significant. We also found that MAP and MRR cannot be optimised simultaneously and this confirmed the theoretical argument about the trade-off between MAP and MRR in [12].

4 Conclusion and Future work

We have proposed a theoretical unified retrieval framework based on the Eliteness Hypothesis. Our initial experiments with a basic ranking function demonstrated that the approach is indeed encouraging. One of the problems with our current estimation is that the EM algorithm does not always converge to the global maximum. In the future, we would like to test the model by estimating the eliteness probabilities with Bayesian methods.

References

1. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statistical Society, Series B*, 1977.
2. S. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Retrieval Science*, 1975.
3. G. Jagadeesh, S. Robertson, and J. Wang. A unified relevance retrieval model by eliteness hypothesis. ArXiv e-prints, `http://arxiv.org/abs/1106.2946`, 2011.
4. M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 1960.
5. J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, 1998.
6. S. Robertson. The unified model revisited. In *Presented at SIGIR 2003 Workshop on Mathematical/Formal Models in Information Retrieval*, 2003.
7. S. Robertson, M. E. Maron, and W. S. Cooper. Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1982.
8. S. Robertson and K. Spark Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 1976.
9. S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 2009.
10. S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter. Probabilistic models of indexing and searching. In *SIGIR*, 1980.
11. S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR*, 1994.
12. J. Wang and J. Zhu. On statistical analysis and optimization of information retrieval effectiveness metrics. In *SIGIR*, 2010.