Ranking Documents Through Stochastic Sampling on Bayesian Network-based Models: A Pilot Study

Xing Tan[1], Jimmy Xiangji Huang[1] and Aijun An[2]
Information Retrieval and Knowledge Management Research Lab
[1]School of Information Technology, [2]Department of Computer Science & Engineering
York University, Toronto, Canada
{xtan, jhuang}@yorku.ca, aan@cse.yorku.ca

ABSTRACT

Using approximate inference techniques, we investigate in this paper the applicability of Bayesian Networks to the problem of ranking a large set of documents. Topology of the network is a bipartite. Network parameters (conditional probability distributions) are determined through an adoption of the weighting scheme tf-idf. Rank of a document with respect to a given query is defined as the corresponding posterior probability, which is estimated through performing Rejection Sampling. Experimental results suggest that performance of the model is *at least* comparable to the baseline ones such as $BM25$. The framework of this model potentially offers new and novel ways in weighting documents. Integrating the model with other ranking algorithms, meanwhile, is expected to bring in performance improvement in document ranking.

Keywords

Info. Retrieval; Bayesian Networks; Stochastic Sampling

1. INTRODUCTION

Probabilistic Graphical Models [10] in the form of Bayesian Networks (BN) [6] are widely used to represent knowledge with uncertainties. In the recent years, computational technologies and tools for BN-based models are becoming increasingly powerful. That being the case, modeling problems in Information Retrieval (IR), in particular for document ranking, as probabilistic inference problems in BN-based models has achieved only limited success to date. Major reasons for such relatively small progress in BN-based approaches for document ranking are, for one, conceptually it is challenging to appropriately identify causalities in document ranking (i.e., deciding network topology) and then to accurately capture the uncertainties (i.e., deciding network parameters) in order to construct a BN model for IR; and two, computationally exact inference algorithms associated with the model is bound to be intractable as practically the size of the BN model in terms of nodes representing both the number of documents and vocabulary size in words, can be easily in a few millions.

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914750

In this paper, we investigate the applicability of BNs to the problem of ranking a large set of documents. We propose a model, which specifically takes into considerations both the appropriateness in the semantics of causalities, and the computational tractability of probabilistic inferences. Topology of the network is a bipartite. Conditional probability distributions are determined through adopting the weighting scheme tf-idf. Experimental results, obtained from working on both computer-generated and standard document sets suggest that performance of the model is *at least* comparable to the baseline ones such as $BM25$ ([2, 11]).

The remainder of this paper is organized as follows. Section 2 presents the background and preliminaries. Section 3 introduces the model. Experimental results are reported and analyzed in Section 4. Section 5 concludes the paper.

2. PRELIMINARIES

In this section, document ranking in IR, and BN, are briefly reviewed.

2.1 Document Ranking in IR

One of the fundamental tasks in IR is document-ranking: Given a set of documents $\mathcal{D}$ such that $\mathcal{D} = \{d_1, \ldots, d_M\}$ and $|\mathcal{D}| = M$, a set of terms $\mathcal{T} = \{t_1, \ldots, t_N\}$ and $|\mathcal{T}| = N$, and a collection of query terms $\vec{q}$ such that $\vec{q} \subset \mathcal{T}$, documents in $\mathcal{D}$ need to be ranked in a complete order according to their respective relevance to $\vec{q}$ (other criteria such as "diversity" might also be considered for ranking). To do this, a typical approach is to define a score function $score(\vec{q}, d_i)$ that returns a numeric score for each document $d_i \in \mathcal{D}$ with respect to $\vec{q}$. Documents can then be ranked on their scores in descending order. The top S elements will be selected to construct the set $\mathcal{S}$, where $|\mathcal{S}| = S$.

Let $tf_{t,d}$ denote *term frequency*, the number of occurrences of term t in document d; df_t denotes *document frequency*, the number of documents in $\mathcal{D}$ that contain the term t; and $idf_t = \log(M/df_t)$ denotes *inverse document frequency*. Summing up on $tf_{t,d} \times idf_t$ for each term $t \in \vec{q}$ with respect to d defines a baseline score function for document ranking: $score(\vec{q}, d) = \sum_{t \in \vec{q}} (tf_{t,d} \times idf_t)$. Definitions for variant score functions of tf-idf such as BM25, can be found in [11]. In addition, *collection frequency* in $\mathcal{D}$ and its subset $\mathcal{S}$, are defined as the total number of occurrences of t in $\mathcal{D}$ and in $\mathcal{S}$, and denoted by cf_t and sf_t, respectively.

2.2 Bayesian Networks

A Bayesian Network is a directed acyclic graph where nodes correspond to random variables [6]. Pairs of nodes in the graph might be connected by a directed edge. For example, given two nodes X and Y in the graph, if X enters Y, it is said that X is a parent of Y. Effect of the par-

ents of a node X in the graph is quantified by a conditional probability distribution $P(X|Parents(X))$.

BNs are often used to carry out probabilistic inference: computing posterior distribution of a set of variables given a set of *evidence variables* – variables whose values are observed or assigned. Consider a simple example of *Baby World*, which consists of four random variables H, T, C, and R, corresponding to the variable facts that the baby is *Hungary*, *Tired*, *Crying*, and *scReaming*, respectively. A BN for this world is shown in Figure 1. After all four conditional probability distributions as listed in the figure are specified, we could compute, for example, the probability that the baby is hungry if we observe that it is crying but not screaming: $P(H \text{ is true} \mid C \text{ is true, and } R \text{ is false})$.

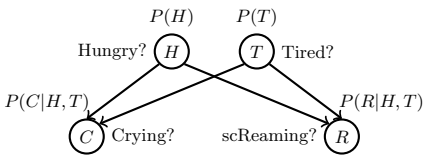

Figure 1: A Bayesian Network: *Baby World*.

3. MODEL

This section presents the model, explains how samplings are performed, and justifies the merits of our model.

3.1 Network Topology and Conditional Probabilities

Suppose a user specifies a set of terms as a query $\vec{q}$ for a set of documents $\mathcal{D}$, a subset $\mathcal{S} \subset \mathcal{D}$ of documents need to be retrieved and ranked in a complete order. To be explained in this section, we formulate this problem into the problems of calculating posterior probability values in a BN-based model where the original query is treated as the observed evidence.

In our model, the probability space is induced accordingly by two sets of random variables $\mathcal{D}$ (document random variables) and $\mathcal{T}$ (term random variables), where $D_i \in \mathcal{D}$ for $1 \leq i \leq M$, and $T_j \in \mathcal{T}$ for $1 \leq j \leq N$. Each document D_i takes two values: $Val(D_i) = \{d_i^1, d_i^0\}$, which represents the values of "D_i selected with respect to a query" (d_i^1) or not (d_i^0); Similarly, $Val(T_j) = \{t_j^1, t_j^0\}$, which represents the values of term "T_j is a query term" (t_j^1) or not (t_j^0). The BN, as shown in Figure 2, is a two-layer directed graph, which contains a node in the top layer for each document variable D_i and a node in the bottom layer for each term variable T_j. In the graph, an edge from D_i to T_j represents that term T_j appears in the document D_i. We assume no edges between document variables in $\mathcal{D}$, and no edges between term variables in $\mathcal{T}$. In addition to the graph, two types of

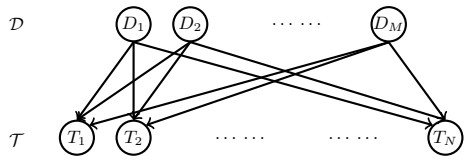

Figure 2: A BN graph for document ranking.

probabilities need to be specified to capture the nature of the dependence of variables: $P(D_i)$, the prior distribution over a document D_i, and $P(T_j|D1, \ldots, D_M)$, the conditional probability distribution of T_j given D_1, D_2, ..., and D_M. In our model, $P(D_i)$ represents the distribution that D_i is selected in $\mathcal{S}$ or not, hence it is reasonable to define, for any document D_i in $\mathcal{D}$, the probability that D_i is eventually selected into S equals the ratio of the size of $\mathcal{S}$ to the size of $\mathcal{D}$, i.e., $P(d_i^1) = S/M$.

The conditional distribution $P(T_j|D1, \ldots, D_M)$ specifies the distribution over the term T_j, which depends on the actual content of $\mathcal{S}$, the set of documents selected. That is to say, specifically, for each subset of $\mathcal{D}$ (totally 2^M of them), we need to specify a distribution for t_j^1, the event that t_j is actually a query term. Since the number of parents of a term in the network is not bounded by a constant, we know that exact inference here has exponential time complexity in worst cases. Nevertheless what we really need is to calculate, for any document variable D_i where $1 \leq i \leq M$, the relevance of D_i to evidence $\tilde{\mathbf{q}}$, i.e., the posterior probability of $P(d_i^1|\tilde{\mathbf{q}})$, the value of which can be estimated through stochastic sampling. For simplicity in explanation, we assume that $\tilde{\mathbf{q}}$ contains only one term $\mathbf{t}$, without loss of generality.

3.2 Estimating Posteriors Through Sampling

General reasoning problems of probabilistic inference in BNs, and even their corresponding approximate versions, are NP-hard ([4], [5]). Due to this computational intractability, one often turns to randomized sampling methods (e.g., Rejection Sampling [12], which is used in our current research) to approximate posterior probabilities. Asymptotically accuracy of these sampling methods would usually be improved as the number of samples increases.

Given a BN and its specified prior distributions for all n variables $\{X_1, X_2, \ldots, X_n\}$, where $X_i \in \mathcal{X}$ for $1 \leq i \leq n$, *forward sampling* samples all nodes in $\mathcal{X}$ in an order consistent with the topological structure of the BN, and the probability of a specific event, written as $(x_1, \ldots, x_n)$[1], generated from forward sampling equals to $\prod_{i=1}^{n} P(x_i|\, parents(X_i))$, which in turn equals to the joint distribution $P(x_1, \ldots, x_n)$. Suppose totally N samples are taken, and the number of occurrences of an event $(x_1, \ldots, x_n)$ equals to $N_{(x_1,\ldots,x_n)}$, then the ratio $N_{(x_1,\ldots,x_n)}/N$ is an approximation to $P(x_1, \ldots, x_n)$. With observation of evidence $\mathbf{e}$ for $\mathcal{E}$, where $\mathcal{E} \subset \mathcal{X}$, the conditional probability $P(X = x|\mathbf{e})$ can be further estimated through *Rejection*: first, all samples that do not match $\mathbf{e}$ are rejected in N, to obtain N_1; second, all samples in N_1 and compatible with $X = x$ are put into N_2; third, N_2/N_1 is an estimate to $P(X = x|\mathbf{e})$.

Consider our model again and suppose there are $\mathcal{P}$ samples where $|\mathcal{P}| = P$. During sampling, we dynamically maintain a vector of counters C, where $|C| = M$. All counters in C are initialized to zero. Our sampling strategy say for the j^{th} sample P^j where $1 \leq j \leq P$: Step 1, each document variable is sampled according to the distribution S/M; Those selected document variables are put into $\mathcal{S}^j$, thus the expected value of $|\mathcal{S}^j|$ is S. Step 2, we accept this sample if and only if the collection frequency of $\mathcal{S}^j$ with respect to term t proportionally exceeds the one of $\mathcal{D}$. Formally, the sample is accepted iff $\frac{sf_t^j}{cf_t} > \frac{S}{M}$. If sample P^j is accepted and $D_i \in \mathcal{S}^j$, $c_i \in C$, which is the corresponding counter for D_i, would be increased by one.

After completion of sampling, the set $\mathcal{P}$ would be mutual-exclusively partitioned into two sets, the set of accepted samples $\mathcal{P}^{accepted}$, and the set of rejected ones $\mathcal{P}^{rejected}$. The vector C would be updated for $|\mathcal{P}^{accepted}|$ times. Values

[1]The term $(x_1, \ldots, x_n)$ is an abbreviation for $(X_1 = x_1, \ldots, X_n = x_n)$.

stored in the counters of C, are actually scores for their corresponding documents with respect to the query term t. The documents can thus be ranked according to their scores.

3.3 Justification of Methodology

In the literature, considerable research in investigating potential linkages between BNs and IR in general, has been reported (most notably [1], [7], [13], [14]). The originality and value of our research contribution lies in the following facts.

Model Semantics. The model defines $2^{(M+N)}$ different states, for different combinations of truth assignments to all random variables in $\mathcal{D} \cup \mathcal{T}$. A state specifies an instance of which random variables are true and which are false. For each state, its joint probability distribution theoretically can be calculated (although computationally it might be impractical). We are only interested in those states where statistically S out of D variables are true, since we only concern about the problem of selecting $\mathcal{S}$ out of $\mathcal{D}$ documents related to a given query.

Causality. In the model, document variables in $\mathcal{D}$ are designed, in consistence with common perceptions, to have direct causal influence on term in $\mathcal{T}$. For example, causal relation "$D_i \rightarrow T_j$" is interpreted as: If D_i is selected to be a member of $\mathcal{S}$ (i.e., $D_i \in \mathcal{S}$) then the term T_i should be of interest to the user.

Scalability. The size of the problem of document ranking in practise is often in the magnitude of a few millions, if not more. A network built-up from these problems is large in size and multiply connected, making it dauntingly challenging to perform exact probabilistic inference. Consequently, it can be seen from the literature that experiments in earlier work (e.g., [7] and [1]), are restricted to cases with maximal a few thousand documents only. The development of BNs however has reached the point where approximate inference algorithms, such as randomized sampling, loopy propagation, or variational approximation, can make a practical difference. We adopt a direct sampling method in this research.

Bipartite Network Structure. The underlying undirected graph of the network is a bipartite: nodes are grouped into two types, only connections between nodes from two different types are allowed. Recently, Bayesian models on bipartite graphs have found their ways to modeling real-world applications in social networks, with appealing properties demonstrated [3]. It remains to be investigated how these results can be utilized into our own framework of BN for IR. Nevertheless, due to this simplicity of topological structure, additional features (e.g., ontological/relational/logical representation and reasoning: see [8] for a mosaic of such proposals) can be incorporated into the current model.

4. EXPERIMENTAL RESULTS

In this section we compare the performance of our proposed model with the ones of tf-idf, $BM25$ and *Golden* (to be explained in Section 4.1). We first work on a set of computer-generated random documents ($\mathcal{D}_R^1$, $\mathcal{D}_R^2$, $\mathcal{D}_R^3$ and $\mathcal{D}_R^4$) and then on $\mathcal{D}_1$ and $\mathcal{D}_2$, which are two subsets of WT2G, a standard TREC test collection.

4.1 Documents Generation

To simplify matters, we assume that all generated documents in $\mathcal{D}_R$ are with same document length, which equals to the size of the vocabulary $\mathcal{T}_R$. For any document $D_r \in \mathcal{D}_R$, occurrences of terms in D_r follow a Normal Distribution $N(\mu, \sigma)$, where the values of mean μ and the standard deviation σ can be adjusted. However all documents in $\mathcal{D}_R$ share the same μ and σ. Note that the smaller the value of σ, the terms cluster more closely to μ. The center of a given document is shifted to a random center before being stored into a vector. To illustrate the idea, an example document with $\mu = 50$, $\sigma = 17$, $|\mathcal{D}_R| = 10000$, and $|\mathcal{T}_R| = 100$, and its variant with shifted center (now, $\mu = 18$) is respectively shown in the left subplot (and the right one) of Figure 3.

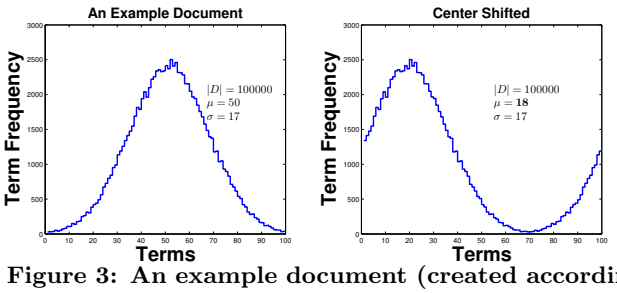

Figure 3: An example document (created according to a Normal Distribution with $\mu = 50$ and $\sigma = 17$), and its variant with the center shifted to 18 ($\mu = 18$).

Accordingly we introduce the baseline ranking method *Golden*. That is, given an input query q and a document d, the $score(q, d)$ is defined as the difference between the center of d, and q. When the value of $score(q, d)$ is smaller (greater), it means document d is more (less) related to the term q. *Golden* is used as one of the three methods for comparisons in Figure 4.

4.2 Experiments

We first work on four sets of randomly generated documents: $\mathcal{D}_R^1$, where $\sigma = 1250$, $\mathcal{D}_R^2$, $\sigma = 1000$, $\mathcal{D}_R^3$, $\sigma = 833$, $\mathcal{D}_R^4$, $\sigma = 416$. Between these sets from $\mathcal{D}_R^1$ up to $\mathcal{D}_R^4$, documents are more closely clustered around their means, thus intuitively more reliable in the sense of IR. Collection size, document length, and vocabulary size, are set to be all equal: $|\mathcal{D}_R| = |\mathcal{T}_R| = 10000$. Experiments related to a specific set is pictorially summarized in the corresponding sub-figure in Figure 4.

Consider Figure 4.a, for example, a term is queried against $\mathcal{D}_R^1$, three different ranking methods, i.e., Golden, BM25, and Rejection, return three different completely ordered sequences of 500 elements (the number 500 is obtained from $|\mathcal{D}_R| \times 0.05$, where 0.05 is the pre-specified ratio, i.e., portion of all documents in $\mathcal{D}_R^1$ need to be ranked). Results of pair-wise comparisons between these three methods are reported in Figure 4.a, where x-axis values indicate sample sizes and y-axis values indicate how many documents out of the 500 ranked ones are actually agreed between two given methods. For example, the red dot pointed by the arrow in Figure 4.a refers to the fact that totally 336 documents are shared by the 500 documents retrieved from applying BM25 on $\mathcal{D}_R^1$, and the 500 elements obtained from performing Rejection Sampling on $\mathcal{D}_R^1$ (with the sample size equaling to 0.1 Million).

Documents in both $\mathcal{D}_1$ and $\mathcal{D}_2$ (Figure 5) are drawn from dataset collection WT2G where $|\mathcal{D}_1| = |\mathcal{D}_2| = 2500$, $|\mathcal{T}_1| = 50961$ and $|\mathcal{T}_2| = 127487$. First 100 elements obtained from three different ranking methods, tf-idf, $BM25$, and *Rejection* are pair-wise compared in Figure 5.

4.3 Brief Discussion

This section draws the major observations from the present experimental study, and discusses some implications.

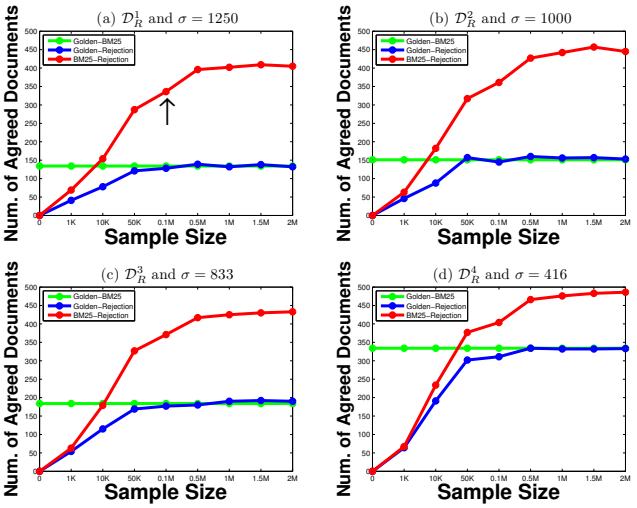

Figure 4: Pair-wise comparisons among the ranking methods Golden, BM25, and Rejection on data-sets with different standard deviations.

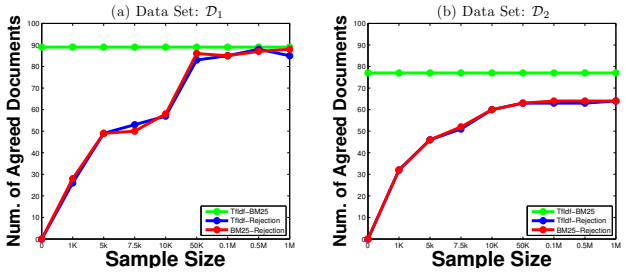

Figure 5: Pair-wise comparisons among the ranking methods tf-idf, BM25, and Rejection on two data-sets (2500 documents each) from WT2G.

For more closely clustered documents, in essence all methods agree more on their rankings (as shown in Figure 4, they agree most on the set $\mathcal{D}_R^4$, but least on $\mathcal{D}_R^1$); Following the same argument, we should claim that $\mathcal{D}_1$ is more clustered than $\mathcal{D}_2$.

In our experimental settings, increasing sample size initially improves the performance sharply, but it tends to be leveling off after sample size is greater than certain value (e.g., 0.5M in the subplots of Figure 4). As the sample size increases, asymptotically the *Rejection* method agrees at least 80% with BM25 for almost all sets except for $\mathcal{D}_2$ (around 60% only). It seems that we should conclude that the proposed BN-based model can achieve competitive performance levels at relatively low cost in sampling.

Since in *Rejection*, estimating posterior probabilities are based on the tf-idf ranking scheme, *Rejection* is a stochastic variant to the tf-idf ranking method. $BM25$, meanwhile, can be deemed as a refinement to tf-idf. Hence, it is not a surprise that $BM25$ agrees largely with *Rejection*. The more intriguing observation is that, with standard data-sets, the two methods disagree at least on 20% of their rankings. The reason as we speculate is either *Rejection* behaves somewhat differently from $BM25$ in practise, or sampling with the current settings can get trapped and do not converge. In order to unravel this intricacy, a further investigation is necessary and desirable.

5. SUMMARY & FUTURE WORK

In this paper, document ranking in IR is transformed into the problems of estimating posterior probabilities through stochastic sampling on a BN designed for IR. Experimental results from this pilot study is quite encouraging in the sense that, with moderate sampling efforts, the model demonstrates its ranking capability comparable to $BM25$. Additionally, graph-based structure and probability-based parameters of the model, together with other considerations in the model, suggest that new and novel weighing schemes for document ranking are conjecturally within a reach.

Among many possible avenues, our direct future research includes 1) further evaluating performance of the model on WT2G, WT10G, and other standard dataset collections; 2) testing on parametric settings other than the current one that is based on term-frequency; 3) testing other sampling strategies (e.g., Gibbs Sampling [10], and the most recent ones [9]) to improve sampling efficiency and performance.

Feedback to this work we have received encourages us to investigate in a broader context the relationships between the proposed BN model and other term weighting models ([15, 16]). With ease, most existing probability theory-based models in IR can actually be derived in this BN-based framework. It is thus rather promising to exploit this framework for a deeper understanding of existing term weighting models, and for the developments of new and better models in Information Retrieval.

6. ACKNOWLEDGEMENTS

We gratefully acknowledge the anonymous reviewers for their insightful comments and suggestions. This research is supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), an NSERC CREATE award in ADERSIM[2] and an ORF-RE (Ontario Research Fund - Research Excellence) award in BRAIN Alliance[3].

7. REFERENCES

[1] S. Acid, L. M. de Campos, J. M. Fernández-Luna, and J. F. Huete. An information retrieval model based on simple Bayesian networks. *Int. J. Intell. Syst.*, 18(2):251–265, 2003.

[2] M. Beaulieu, M. Gatford, X. Huang, S. Robertson, S. Walker, and P. Williams. Okapi at TREC-5. In *Proc. of TREC*, pages 143–166, 1996.

[3] F. Caron. Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems 25*, pages 2051–2059. Curran Associates, Inc., 2012.

[4] G. F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2):393 – 405, 1990.

[5] P. Dagum and M. Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60(1):141 – 153, 1993.

[6] P. A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[7] R. Fung and B. Del Favero. Applying Bayesian networks to information retrieval. *Commun. ACM*, 38(3):42–ff., Mar. 1995.

[8] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[9] K. Kandasamy, J. G. Schneider, and B. Póczos. Bayesian active learning for posterior estimation - IJCAI-15 distinguished paper. In *Proc. of the 24th IJCAI*, pages 3605–3611, 2015.

[10] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[11] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[12] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[13] H. Turtle and W. B. Croft. Inference networks for document retrieval. In *Proc. of the 13th ACM SIGIR*, pages 1–24, New York, NY, USA, 1990.

[14] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9(3):187–222, July 1991.

[15] J. Zhao, J. X. Huang, and B. He. CRTER: Using Cross Terms to Enhance Probabilistic IR. In *Proc. of the 34th ACM SIGIR*, pages 155–164, 2011.

[16] J. Zhao, J. X. Huang, and Z. Ye. Modeling Term Associations for Probabilistic Information Retrieval. *ACM Trans. Inf. Syst.*, 32(2):1–47, 2014.

[2] http://www.yorku.ca/adersim

[3] http://brainalliance.ca