A Tri-Valued Belief Network Model for Information Retrieval

December 2001

Fernando Das-Neves
Computer Science Dept.
Virgina Polytechnic Institute and State University
Blacksburg, VA 24060

1. IR models at Combining Evidence

Graphical Models [Pearl88] [Buntine94] [Jensen2001] allow for greater flexibility at modeling relations among different sources, when compared with other methods like LSI, or standard vector models, because each source can be modeled independently. Graphical models rely on a subjective instead of frequentist understanding of probabilities. Such probabilities do not necessarily represent relative frequencies (although they are sometimes calculated like they were), but instead they represent the degree of certainty, belief or support about a certain feature (a term, a link, etc.), and about one feature given another feature (conditional probabilities). The most common type of graphical model applied to IR is Bayesian Networks. The first successful application of Bayesian networks to IR is the Inference Network Model in [Fung95], [Turtle90]. Belief Networks [Ribeiro-Neto96] [Silva2000] are an alternative approach to Inference Networks when explicitly modeling an information retrieval system. While similar in expressive power to inference networks, belief networks can express any inference network used to retrieve documents by content similarity, while the opposite is not necessarily true. The key difference is in the modeling of $p(d_j|t)$ (probability of a document given a set of terms or concepts) in belief networks, as opposed to $p(t|d_j)$ used in Bayesian networks. Since in a Bayesian network (both of the inference and belief type), instantiating d_i makes $p(t_1|d_j)$ and $p(t_2|d_j)$ defined and mutually independent, then $p(t|d_j)$ can be calculated from the independent probabilities $p(d_j|t_i)$ in a belief network, and so belief networks can reproduce the ranking generated by an inference network. Generating the ranking produced by belief network from an inference network is not always possible, as a belief network can reproduce the cosine similarity ranking, while this is not possible for an inference network (see [Ribeiro96] for details). Figure 4 illustrates the differences in the probability structure between the two networks (note the direction of the arrows between the terms and the documents):

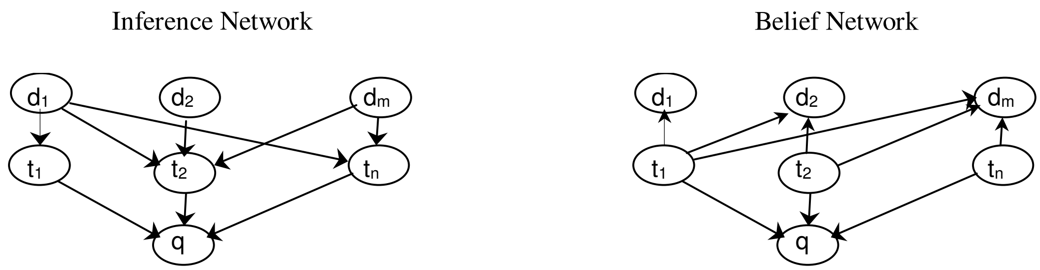

Figure 4. Examples of Inference and Belief Networks to calculate relevance of a document given a query.

In these graphs, an arrow between two nodes represent the conditional probability of the variable pointed by the arrowhead, given the value of variable at the tail of the arrow. Variables not connected by arrows are assumed to be independent. While these networks look almost identical, the conditional probability structure is different, and it has two important consequences:

1) The “hammock" structure between a document and a query can be made to represent a vector dot-product, with $p(d_i|t)$ and $p(q|t)$ the two vectors of dimension n to multiply, with $p(d_i|t_j)$ and $p(q|t_j)$ being the individual components of the vectors. From this point of view, belief networks constitute a tool to combine the best of vector and probabilistic operations.
2) Belief networks can calculate any ordering calculated by an inference network, including the traditional cosine similarity used in SMART, while the opposite is not true. An inference network, on the other hand, ranks the documents by calculating $p(q|d)$ using the chain rule: for the set U of variables in the model $P(U) = \prod p(\text{var} \mid parents(\text{var}))$. Inference networks can model a structure that calculates a ranking similar to the cosine similarity, but the calculation has an extra term that is document-dependent, and so for Bayesian networks cannot replicate the ranking given by cosine similarity.

It is worth noting that we are showing here only the basic networks. It is possible to include relevance feedback as part of the network as new nodes, or relations between terms like synonyms as new nodes connecting related terms (as long as they do not introduce cycles). New nodes also arise from other sources of information, as clusters, or hubs and authorities like in CLEVER [Kleinberg98]. [Turtle90], [Haines93] and [Silva2000] provide more information on these extensions of the basic network model.

Our retrieval scheme, explained in detail below, allows not only for the specification of document features that are important or irrelevant to the user need, but also allows for a neutral interpretation, letting the user say “I don’t know”. Document features that do not explicitly appear in the query formulation are taken as “undetermined” by the system. Document content is treated as usual.

2. How we perform Retrieval

Our model for retrieval is a belief network combining multiple sources of evidence. A belief network let us combine probabilities and vector operations, allowing us a fast calculation of similarity values even for big collections.

Our belief network is formally described as follows:

2.1 Notation

- An uppercase letter like A letter represents a set.
- One or more lowercase letters without a subindex, (like *da*, for example), are subsets of other sets (context given where is used).
- A lowercase letter with a subindex, like a_i , represents the i^{th} element of the set named with the same letter but in uppercase.
- $|A|$ is the number of elements of set A.
- $\vec{A}$ is the vector representation of the set A, given some ordering of the set elements, and $\vec{a}_i$ is the i^{th} component of $\vec{A}$. Note that vectors and sets are equivalent.

- For a set of variables $x_{1\ldots}x_n$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

2.2 Variables and their Relationships:

2.2.1 Structure.

The definition of the belief network is based on the basic belief network in [Ribeiro96]. In our case, there is one belief network per document d with the following structure:

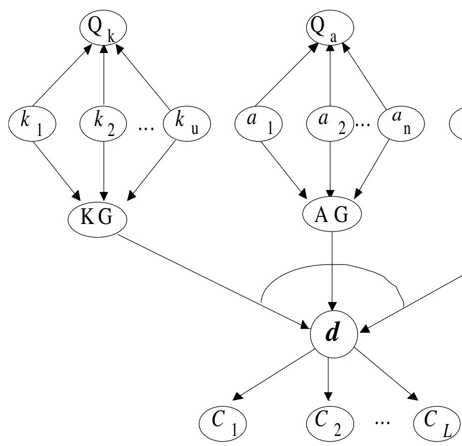

Figure 2. Belief Network to calculate and combine evidence from terms, authors and references.

where

- d is a document,
- D is the set of all documents in the collection,
- Q_k; Q_a; Q_r represent the query information about document content, authors and references, respectively,
- KG; AG;RG represent the set of Keywords (Keyword Group), Authors and
- References, respectively,
- $\vec{q}_{ki}$, $\vec{q}_{a_k}$, $\vec{q}_{r_n}$ are the vector representations of $q_{ki} \in Q_k$, $\vec{q}_{a_k} \in Q_a$, $\vec{q}_{r_n} \in Q_r$, for $1 \le i \le |KG|, 1 \le j \le |AG|, 1 \le n \le |RG|$,
- k_i, a_i; r_i represent a particular author, a keyword or a reference,
- C_l represents a document cluster, for $1 \le l \le |D|$.

By the structure of the belief network, we treat content-based and structural information the same way. Variables k_i , a_i and r_i represent a link (although not explicit) between all documents that contain the information represented by the variable, which must appear in at least one document for the variable to belong to the bayesian network.

2.2.2 Variables.

The following table describes the domain (possible set of values) of the variables described above:

Variable	Domain
Q_k ; Q_a ; Q_r	{*1, 0*}
$\vec{q}_{k\,i}$, $\vec{q}_{a_k}$, $\vec{q}_{r_n}$	{*yes, no, unknown*}
k_i, a_i, r_i	{*yes, no*}
KG, AG, RG	{*relevant, irrelevant*}
d	{*relevant, irrelevant*}
C_l	{*1, 0*}

Table 2. Variables of the Belief Network in Figure 2, and their possible values.

2.3 Interpretation of variable values.

For a certain document *d*; k_i = *yes* iff the keyword or concept represented by k_i is relevant for *d*, otherwise k_i is considered non-relevant for *d* and k_i = *no*. In the same way and for the same document *d*, a_i = *yes* or r_i = *yes* if the corresponding author or reference appears in *d*. We call the sets *KG*; *AG* and *RG* sources of evidence of *d*.

A query *Q* from the user is represented as three independent and disjoint query subsets Q_k; Q_a ; Q_r, so Q = $Q_k \bigcup Q_a \bigcup Q_r$ Q, and $q_{k\,i} \in Q_k$, $\vec{q}_{a_i} \in Q_a$, $\vec{q}_{r_i} \in Q_r$ .If the query formulation included the keyword k_i then $q_{k\,i} = yes$, otherwise $q_{k\,i} = unknown$ **unless the user has explicitly said** that $q_{k\,i}$ is not relevant for the query subset Q_k, case in which then . $q_{k\,i} = no$. For example, if Q_k = {k_1; k_4 ; *not*(k_{10})}, then $q_{k1} = q_{k_4} = yes$, $q_{k10} = no$ q k10 = no, and all other $q_{k\,j} = unknown$. Under this interpretation, absence of evidence about relevance of a keyword, author or reference in the query formulation does not constitute proof of its irrelevance.

The same definitions also apply for all the $q_{a\,i}$ and $q_{r\,i}$.

Any of query subsets Q_k ; Q_a and Q_r is relevant if the probability of at least one of the elements of the query subset is non-zero. That is, if we call *G* one of the query subsets of *Q* and there is some $v_i \in G$ such that $p(G|v_i) \neq 0$, then *G* is relevant. A document *d* is relevant given a query *Q* if at least one of Q_k; Q_a; Q_r , is relevant.

A cluster C_l = 1 if at least one document belongs to C_l, and $p(C_l = 1|d)$ is the probability of the information in cluster C_l being related to the content of document *d*.

2.4 Definition of Probabilities:

The advantage of this type of belief network is that every hammock structure can be interpreted as a vector dot-product, and so it let us combine vector operations with probabilistic operations.

When combining evidence for all the sources, we want that

- Absence of evidence from one source does not affect other sources
- Evidence from more than one source is more effective that evidence from only one source
- If a source of evidence uniquely identifies a document, all other sources are irrelevant.

A noisy-or [Good61] of the sources satisfies the above conditions, under the added assumption that all sources are independent:

$$p(d = relevant \mid KG = relevant; AG = relevant; RG = relevant) = \\ 1 - [(1 - wk(p(d \mid KG))) \times (1 - wa(p(d \mid AG))) \times (1 - wr(p(d \mid RG)))] \quad [1]$$

Here *wk*, *wa*, and *wr* weight-adjusting functions, that transform each of the probabilities according to some measure of the importance of the sources. The functions *wk*, *wa*, and *wr* must be monotonically increasing and between 0 and 1.

For the sake of deriving some of the probabilities, and to simplify the explanation, we are going to use *AG* as the source of evidence, an a_i and as the particular instantiation of that source in a particular document. Because of the similar structure, the same rationale can be applied also to *KG* an *RG*, unless otherwise noticed.

Let us define the sets A = { set of all authors in all documents }, and the set a to be a subset of authors such that $a \subset A$, and vectors $\vec{d}_a$ and $\vec{dn}_a$ such that

$\vec{d}_{a_i} \neq 0$ if a_i is an author of d, and 0 otherwise, and

$\vec{dn}_{a_i} \neq 0$ if a_i is not an author of d, and 0 otherwise.

By the fundamental rule of probabilities we have $p(AG \mid Qa) = \dfrac{p(AG \cap Q_a)}{p(Q_a)}$, and

$$p(AG \cap Q_a) = \sum_{a_i \in A} p(AG \mid a_i)\, p(Q_a \mid a_i)\, p(a_i) \quad [2]$$

(see the original Belief Network paper [Ribeiro96] for the rationale of this equation).

If we define $p(d \mid AG) = p(AG \mid Q_a)$, then given $p(AG \mid a_i)$; $p(Q_a \mid a_i)$, and $p(a_i)$ we can calculate equation [1].

Let

$x_i = p(AG|a_i = yes)$, $\quad 1 \leq i \leq |a|$
$y_j = p(AG|a_j = no)$ $\quad 1 \leq j \leq |A|-|a|$
$z_k = p(Qa|a_k)$ $\quad 1 \leq k \leq |A|$

x_i is the probability that the author a_i is one of the authors of document d. y_i is the probability that author i is not one of the authors of document d. z_i is the probability that $q_{a_k} = a_k$. *Being 0 otherwise.* We ignore any author that is not an author of any document in the collection.

For the case were for all $q_{a_i} \in Q_a$ each q_{a_i} either *yes* or *no* (the set of authors of a document is completely specified in the query, both the ones that are authors and the ones that are not), we desire a complete exact match between Q_a and A to be the *idf(a)*, the inverse of the number of documents that have the set a and only a as authors. The reason for this is that for an exact specification we want the probability to be proportional to the number of those cases in the document collection.

To say this formally, for the case $\vec{Q}_a = \vec{d}_a + \vec{dn}_a$ (that is, $\vec{Q}_a$ is all 1s), we want, given that $p(a_i)$ is constant for all documents and all authors (a reasonable and common assumption for a-priori author probabilities),

$$\left(\sum_{i=1}^{|A|} x_i z_i + \sum_{j=1}^{|A|} y_j z_j \right) p(a_i) = idf(a)$$

This we can approximate by

$$\left(\sum_{i=1}^{|A|} x_i z_i + \sum_{j=1}^{|A|} y_j z_j \right) = idf(a) \qquad [3]$$

Since $p(a_i)$ is a constant that affects all document in the same way.

Please note that for the same index i., x_i and y_i are never 1 simultaneously.

Note also that $\sum_{i=1}^{|A|} x_i = |da|$ and $\sum_{i=j}^{|A|} y_j = |A| - |dna|$

2.5 Boolean Approach.

One way to give values to x_i, y_i and z_i is

$x_i = 1/|A|$ iff $a_i = yes$, $y_j = 1/|A|$ iff $a_j =$no, and $z_k = 1$ iff $Q_{ak} = a_k$. That is, the probability of an author being relevant or non-relevant is constant.

Then, $\left(\sum_{i=1}^{|A|} x_i z_i + \sum_{j=1}^{|A|} y_j z_j\right) = \sum_{i=1}^{|A|} g_i z_i$, where $g_i = 1/|A|$.

For the case $(\forall i)(p(q_{a_i} \mid a_i) = 1$ (that is, all q_{a_i} matched the corresponding a_i),

we have that [3] then becomes $\sum_{i=1}^{|A|} g_i z_i = \sum_{i=1}^{|A|} g_i = |A| \cdot \frac{1}{|A|} \cdot idf(a) = idf(a)$ as we wanted. Also, for the case if only positive information (that is, only authors who may be authors of d are specified in Q_a), equation [3] becomes $sim(Q_a, d_a) = \frac{|d_a \cap Q_a|}{|A|}$, which is the intersection of the boolean sets Qa and d_a.(Q_a is Boolean in this case because for positive information only, each component has either the value *yes* or *unknown*).

2.6 Another Approach.

However, this is not the only possible definition of probabilities. If we define x_i and z_i then it is also possible to deduce the formulation for y_i , subject to the constraint that the resulting formulation is a probability. Again, for the case $\vec{Q}_a = \vec{d}_a + \vec{dn}_a$ (another way of saying the query included all authors, and matched exactly the set of all authors of d), we have that all $z_i \neq 0$. If we define $p(q_{a_i} = yes \mid a_i = yes) = p(q_{a_i} = no \mid a_i = no)$ = some constant z_c for all i , then we want

$$z_c\left(\sum_{i=1}^{|A|} x_i + \sum_{j=1}^{|A|} y_j\right) = z_c\left(\sum_{x_i \in da} x_i + \sum_{y_j \in dna1}^{|} y_j\right) = idf(a)$$

and therefore

$$|da|\,\bar{x} + (|A| - |da|)\bar{y} = idf(a)\frac{1}{z_c} \qquad [4]$$

Now we need to define x_i. A possible definition is

$$x_i = \frac{idf(\sup erset_{da})}{|da|}, \text{ so that } \sum_{i=1}^{|da|} x_i = idf(\sup erset_{da})$$

where

$$idf(\sup erset_{da}) = \frac{1}{number\ of\ document\ that\ have\ at\ least\ the\ all\ the\ authors\ in\ set\ da}$$

this definition is useful because it relates the probability of a document having a set of authors to the size of the set of document authored by all those authors, with possibly other authors. Note also that $0 \le idf(superset_{da}) \le idf(da)$.

With these definitions of $\bar{x}$ and z_k we can calculate $\bar{y} = \frac{\frac{idf(da)}{z_c} - |da|\bar{x}}{|A|-|da|}$ therefore $y_i = \frac{\frac{idf(da)}{z_c} - |da|\bar{x}}{(|A|-|da|)^2}$.

since $0 \le \bar{x}, \bar{y} \le 1$, for these definitions to be valid they need to satisfy these two conditions:

1) $\frac{idf(da)}{z_c} - |da|\bar{x} \ge 0$, and
2) $\frac{idf(da)}{z_c} - |da|\bar{x} \le |A| - |da|$

Condition 1 implies $idf(\sup erset_{da}) = \sum x_i \le \frac{idf(da)}{z_c}$. Since $0 \le z_c \le 1$ it must be that $df(\sup erset_{da}) \le \frac{idf(da)}{z_c}$, which is only possible to guarantee when $z_c = 1$. Therefore we define $z_k = p(q_{a_i} = yes|\ a_i = yes) = p(q_{a_i} = no|\ a_i = no) = 1$.

With these values for z_k, to probe that the definition of x_i satisfies condition 2 is to prove that $\sum x_i \ge -|A| + |da| + idf(da)$. We have four possible cases:

Case 1) $|da| = 1$ and $idf(da) = 1/|D|$

Here $\sum x_i \ge -|A| + |da| + idf(da) = -|A| + |1| + 1/|D|$. Since $idf(da) = 1/|D|$ and because $1/|D| \le idf(superset_{da}) \le idf(da)$, it must be that $idf(superset_{da}) = idf(da) = 1/|D|$. Since $-|A|+1 \le 0$, it follows that $\sum x_i = 1/|D| \ge -|A| + |1| + 1/|D|$.

Case 2) |*da*| = |*A*| and *idf(da) = 1/|D|*

Following the reasoning in case 1, it must be that $idf(superset_{da}) = idf(da) = 1/|D|$. Furthermore in this case $-|A| + |da| = 0$, then $\sum x_i = \frac{1}{|D|} \geq \frac{1}{|D|}$.

Case 3) |*da*| = |*A*| and *idf(da) = 1*

Now we need to prove $\sum x_i \geq -|A| + |da| + idf(da) = idf(da)$. Since $|da| = |A|$ means all the authors are included in *da*, and $idf(da) = 1$ means there is only one document matching, it follows that there must be only one document in the collection. Therefore $|D| = 1$, and since *da* cannot be empty (we have to be matching at least one author), the only possible answer is $idf(superset_{da}) = idf(a)$ as needed.

Case 4) |*da*| = *1* and *idf(da) = 1*

$\sum x_i \geq -|A| + |da| + idf(da)$ then $\sum x_i \geq -|A| + 1 + idf(da) = -|A| + 2$. We need to look at two subcases, namely $|A| = 1$, and $|A| > 1$.

If $|A|=1$, then $idf(a)=idf(superset_{da})$, because then the same author is the only author of all documents, and condition 2 becomes $1 = \sum x_i \geq -|A| + 1 + idf(da) = idf(da)$.

If $|A|>1$ then condition 2, for case 4 becomes $\sum x_i \geq -|A| + 1 + idf(da) = 0$, which is always true.

Therefore we have proven case 4.

This is all we need to calculate $p(d|AG)$. For all structural information, as References, Journal where the document was published, or Authors, all of the above applies. For content information (set *KG*), we follow the same approach as in [Baeza99] and we define

$p(KG|k_i) = \frac{\vec{d}_{i,j}}{|\vec{d}|}$, where $\vec{d}_{i,j} = idf(k_j) \times tf(k_j)$ if term k_j=*yes* in document d_i, or 0 otherwise, and

$p(Q_k|k_i) = \frac{\vec{q}_{k_j}}{|\vec{Q}_k|}$, where $\vec{q}_{k_j}$ = weight of term kj in query Qk, iff $\vec{q}_{k_j} = k_j$ in *d, or 0* otherwise.

These definitions of $p(KG|k_j)$ and $p(Qk|k_j)$, when applied to equation [2], make the result of equation [2] to be the cosine similarity between the query and the document, times a constant $p(k_i)$.

2.7 Cost of calculating *p(d|Q)*.

Under the above definition of the probabilities in the belief network, it is easy to see that if values like *idf(da)* and $idf(superset_{da})$ are precomputed while the collection is indexed, the cost of calculating $p(d|Q)$ is *O(total number of authors + number of Keywords + number of references)* for each document *d* and query *Q*, since the cost of calculating each probability becomes constant, and calculating the importance of a source is calculating a set of dot-products.

2.8 Document/Cluster probability.

For two documents d_1 and d_2, we define

$$p(d_1|d_2\ ) = p(d_1|Q_k = KG\ of\ d_2\ ;Q_a = AG\ of\ d_2\ ;Q_r = RG\ of\ d_2\ )$$

If for a moment we assume we know $p(d_i|Cn\ )$; then the probability of cluster *n* given document *i*, is $p(C_n|d_i\ ) = \dfrac{p(d_i|C_n\ )p(C_n)}{p(d_i)}$ by Bayes Rule.

If we assume the existence of $|D|$ clusters (as many clusters as documents), and a-priori probabilities $p(C_i) = p(d_i)$ constant for all $1 \le i \le |D|$, then $p(C_n|d_i) = p(d_i|C_n)$, which we can define in many ways, for example $p(d_i \mid C_n) = \frac{1}{|C_n|} \sum_{d_j \in C_n} p(d_i \mid d_j)$ (average similarity between the document and all other documents for which $p(d_j|C_n) \neq 0$: To avoid the problem of all documents depending on all the other documents' probabilities, at the beginning we can initialize $p(d_i|C_n) = 1$ iff $i = n$; and 0 otherwise (at the beginning, each document belongs to its own cluster, which is consistent with the initial state of many clustering algorithms).

3. REFERENCES

[Pearl88] Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[Buntine94] Buntine, W. *Operations for learning with graphical models*. Journal of Artificial Intelligence Research, 1994.

[Jensen2001] Jensen, F. *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[Fung95] Fung, R., and Favero, B. *Applying Bayesian Networks to Information Retrieval*. Communications of the ACM, Vol 38, No 3, 1995.

[Turtle90] Turtle, H., Croft, W. *Inference networks for document retrieval*. In Proceedings of SIGIR'90, pp.1-24, 1990.

[Silva2000] Silva, I., Ribeiro-Neto, B., Calado, P., Moura, E., Ziviani, N. *Link-based and content-based evidential information in a belief network model*. In Proceedings of SIGIR 2000, pp. 96-103, 2000.

[Ribeiro-Neto96] Ribeiro-Neto, B., Muntz, R. *A belief network for IR*. In Proceedings of SIGIR'96. pp. 253-260, 1996.

[Haines93] Haines, D., Croft, B. *Relevance feedback and inference networks*. In Proceedings of SIGIR'93, pp. 2-11, 1993.

[Kleinberg98] Kleinberg, J. *Authoritative Sources in a Hyperlinked Environment.* In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp 668-677, 1998.

[Baeza99] Baeza-Yates, R, Ribeiro-Neto, B. Modern Information Retrieval. ACM Press, 1999.