

Chapter 24

Holistic Query Expansion using Graphical Models

Daniel Mahler
Cycorp, Inc.
3721 Executive Center Dr, suite 100
Austin, TX 78731
mahler@cyc.com

1. Introduction

In this paper we present a method for answering relationship questions, as posed for example in the spring of 2003 evaluation exercise of the AQUAINT¹ program, which has funded this research.

The goal of the exercise was to provide answers to questions requesting an account of the relationship between two or more entities. No restriction on the format of the answer was imposed, except that it should not consist of entire documents and that it should list document identifiers for the passages from which the answer was drawn or generated.

2. Relation Questions

An example question from the relationship exercise and the answer provided by our system, QUIRK, is presented in figure 1.

Question: What does Soviet Cosmonaut Valentina Tereshkova (Vladimirovna) and U.S. Astronaut Sally Ride have in common?

Connecting terms: **first space shuttle woman**

Answer:

APW19990618.0223.11 In 1963, Soviet cosmonaut Valentina Tereshkova

¹ See <http://www.ic-arda.org/InfoExploit/aquaint/>. Participants in this program have access to two corpora a newswire corpus spanning 1996-2000 and a subset of the Center for Nonproliferation Studies abstracts. The AQUAINT evaluations are carried out against these corpora. They are referred to as the AQUAINT and CNS corpus respectively.

returned to Earth after spending nearly three days as the **first woman in space**.

APW19990115.0263.10 In 1978, NASA named 35 candidates to fly on the **space shuttle**, including Sally K. Ride, who became America's **first woman in space**, and Guion S. Bluford Jr., who became America's **first black astronaut in space**.

NYT19990723.0009.1 CAPE CANAVERAL, Fla. After a year of technical delays and two frustrating waits on the launching pad, the **space shuttle** Columbia blasted off from Cape Canaveral early Friday morning carrying the \$1.5 billion Chandra X-ray Observatory, and vaulting its commander, Col. Eileen M. Collins of the Air Force into the ranks of aviation pioneers like Amelia Earhart and the Soviet cosmonaut Valentina Tereshkova, the **first woman in space**.

APW19981020.1367.6 June 18, 1983: Sally Ride becomes **first American woman in space**.

APW19990521.0292.6 Anger in **space**, by astronauts and cosmonauts, has been common since early in the manned **space** program.

APW19990525.0041.29 Former astronaut Sally K. Ride is 48.

Figure 1: A relationship question

What makes this question-answer pair interesting is the fact that no document in the target corpus mentions both Valentina Tereshkova and Sally Ride. This means that the passages that collectively elucidate the relationship between the two must be retrieved from the corpus by

- first finding terms (which we will call *connecting terms*) that are *not* present in the question but which describe (or are evocative of) the relationship between the terms in the question, in this case **first space shuttle** and **woman**, and then
- retrieving relevant passages for the original query expanded with these new terms.

Identifying such *connecting terms* from a corpus is the key aspect of the technique we present in this paper. This technique differs in important respects from other well established query expansion techniques [19,10,2,6] as the following examples are designed to show.²

² Unlike the example in figure 1, these examples are hypothetical and designed for pedagogical purposes. It is not claimed that any system exactly replicates any given example. They are offered in place for the actual output of the QUIRK system to avoid confusing the intuitions we wish to convey to the reader with the imperfections of our

Ideal examples of *connecting terms* for pairs of query terms are:

- France, Germany → Europe country EU
- Rushdie, Khomeini → fatwa, blasphemy
- cats, dogs → pets domestic mammals chase fight

In contrast, based on our understanding of the algorithms, we would expect most other approaches to automatic query expansion to favor terms strongly related to *individual* query terms independently of other terms in the query. Hypothetical examples could be

- France, Germany → French, Paris, German, Berlin
- Rushdie, Khomeini → Salman, Ayatollah
- cats, dogs → Poodle, Retriever, Siamese, Manx

where *Paris* is only related to *France* and *Berlin* is only related to *Germany*. This style of expansion has the potential to lose the focus of the query. For example expanding a query as in

- dogs, training → Poodle, Retriever, Terrier, Setter, jogging, gym, weights

would be likely to cause a search engine to retrieve a mixture of documents about dog breeds and about sports training rather than documents on dog training, which is presumably the user's intent. We refer to this phenomenon as the *outweighing* or *overpowering* of the query by the expansion terms. Identifying *connecting terms* such as

- dogs, training → obedience, sit, heel, leash, reward

cannot always be achieved by considering the query terms in isolation.

2.1 Answer Presentation

The most effective presentation of answers to relationship questions is an issue in itself. Relationship and cause and effect questions differ significantly from factoid questions used in TREC Q&A track competitions, where the goal is to identify a single phrase that gives the answer. Often the relationship is not

current system. For example, in figure 1, **shuttle** is actually returned by our system as a *connecting term* even if it is in fact irrelevant to the query.

described in the corpus in a single concise phrase. Such a description may not be appropriate because of subtlety/complexity of the relationship, as several questions in sample in section 5 should illustrate. Even in cases where the *connecting terms* have been correctly identified, by themselves they usually do not constitute a useful answer, because, unless one already knows the answer, it is not obvious how the *connecting terms* account for the relationship between the query terms³. Even in patent cases such as

- Kennedy, Oswald → kill

if one were ignorant of modern history it would be impossible to rely on the connecting term **kill** to decide between the following interpretations:

- Kennedy killed Oswald;
- Oswald killed Kennedy;
- Kennedy and Oswald both killed someone;
- Kennedy and Oswald were both killed by someone;
- ...

Thus, we believe the most effective use of the discovered *connecting terms* is to expand the original query to find passages which mention the query terms together with the connecting terms. Such passages should then be presented to the user with the terms suitably highlighted, as in figure 1.

3. Identifying Connecting Terms

In this section we consider possible algorithms, which, given a corpus and a query, determine expansion terms that would match our *informal* notion of *connecting terms*. These algorithms then define a family of different *formal* notions of *connecting terms*. We adopt the standard statistical IR assumption that inter-term relevance is reflected in the co-occurrence statistics over the corpus of interest. Applying this principle to the notion of *connecting terms* means that the *connecting terms* should lie *between* the query terms according to some co-occurrence based distance or similarity measure (see section 3.2) on terms [17]. This requires in turn defining what it means for a term *C* to be between two terms *A* and *B* with respect to the given measure. Thus, there are three parameters to making the original *informal* notion concrete:

³ This is an important point to which we return in section 5.

1. the raw co-occurrence data itself;
2. the co-occurrence measure;
3. a notion of an *optimal connection*, for a given set of seed nodes in a weighted graph.⁴

Algorithm 1 (General form of the connecting terms expansion algorithm.)

For some choice of:

- a Probabilistic⁵ Information Retrieval engine E ;
 - a number N of documents to be retrieved;
 - a number of relevant sentences R ;
 - a similarity measure M ;
 - a number W of content words to consider;
 - a similarity measure K ;
 - an algorithm C for retrieving nodes from a graph given a seed set of nodes;
 - a second number J of documents to be retrieved;
1. Compute query terms $\{q_1, \dots, q_n\}$ from a query Q ; in a Q&A exercise this may involve stripping question words and stop words.
 2. Submit $\{q_1, \dots, q_n\}$ to a search engine E and retrieve the top N documents;
 3. Split documents into sentences and remove stop words from sentences;
 4. Select the R sentences $\{s_1, \dots, s_R\}$ that are most relevant to Q by some measure of similarity M between $\{q_1, \dots, q_n\}$ and s_i ;
 5. Build word-sentence matrix for the W words $\{w_1, \dots, w_W\}$ that occur in the greatest number of sentences in $\{s_1, \dots, s_R\}$;
 6. Use similarity measure K to build a graph representation G of the pairwise similarity between any two w_i, w_j in $\{w_1, \dots, w_W\}$;
 7. Use method C to compute a set $\{c_1, \dots, c_m\}$ of connecting terms for $\{q_1, \dots, q_n\}$ from the graph G ;
 8. Submit $\{q_1, \dots, q_n\} \dot{\cup} \{c_1, \dots, c_m\}$ to the search engine E and retrieve the top J documents;
 9. Split documents into sentences and cluster them by agglomerative clustering [8];

⁴ This would be a connected subgraph of the initial graph that contains all the seed nodes and optimizes some property, as a shortest path or a maximum spanning tree does. Any nodes in this subgraph, but not in the seed set, can be considered to be *between* the seed nodes in the initial graph.

⁵ A probabilistic retrieval model enables us to obtain good partial matches for a set of terms without us having to construct complex boolean query expressions.

10. Return a representative sentence from each cluster;

These parameters define a very large design space that results from instantiating the algorithm schema in algorithm 1. The steps that distinguish the *connecting terms* expansion algorithm from other forms of query expansion are highlighted.

Information Retrieval Engine (E)	Lemur
Number of documents retrieved (N)	25
Number of relevant sentences (R)	1/7 of all sentences in the N documents
Similarity measure (M)	overlap
Number of content words (W)	400
Similarity measure (K)	<i>weight of evidence</i> [9,7]
Algorithm for extracting <i>connecting terms</i> from graph (C)	smallest connecting subgraph of maximum spanning tree of G that contains $\{q_1, \dots, q_n\}$
Number of documents retrieved (J)	50

Figure 2: Current parameter values for the QUIRK *connecting terms* feedback algorithm.

Figure 2 lists the particular values currently used by the QUIRK system. Below we discuss our choice for the values of the most significant parameters and alternatives that would be worth exploring.

3.1 The Definition Of Connecting Node

In order to compute the list of *connecting terms*, we cast the problem in terms of graphs. We construct a graph G that has as nodes the union of the original query terms and a set of candidate expansion terms. The procedure by means of which the set of candidate expansion terms is selected is one of the three parameters of the *connecting terms* expansion algorithm schema and is explained in section 3.3 below. G is fully connected and the edge between nodes A and B is assigned as weight the similarity between A and B as gleaned from the corpus co-occurrence statistics, according to some measure (step 6 in algorithm 3.). In the QUIRK system we define the *connecting terms* for a query as all the nodes in G that can be found on the smallest connected graph which

1. contains all query terms; and
2. is a subgraph of G 's maximum spanning tree (e.g.[14]).

The method for extracting this set corresponds to parameter C in step 7 of the *connecting terms* expansion algorithm schema.

Figure 3 is meant to depict the maximum spanning tree of some graph. The set of nodes highlighted in grey represents the smallest connected subgraph that contains $Q1$ and $Q2$. If $Q1$ and $Q2$ represent terms from a query (as *Tereshkova* and *Ride* might be from the example in figure 1) then $C1$, $C2$ and $C3$ represent *connecting terms* for it, as **first**, **space** and **woman** might be in the same example.

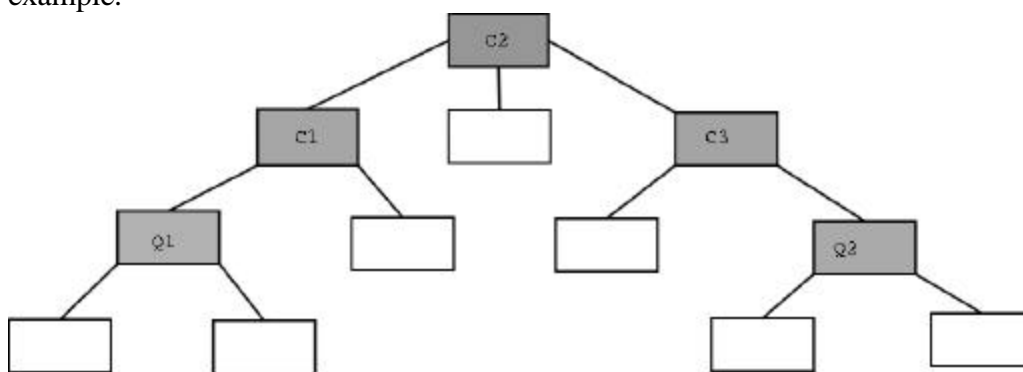


Figure 3: An example of *connecting terms* selection

While the maximum spanning tree method proved to be satisfactory in the selection of *connecting terms*, one could readily think of plausible alternatives. Among them, other network algorithms such as *maxflow* and *mincut* algorithms [14], *Bayesian Propagation* as described in [12,11] or algorithms based on spreading activation such as [13,4]. Choosing one of these other methods would result in a different definition of what it is for a node to be a *connecting term* for a query, and so in a different instantiation of the *connecting terms* expansion algorithm schema.

3.2 Co-occurrence Measures

A second parameter of the *connecting terms* expansion algorithm schema is the particular choice of the similarity measure between two terms that appear in the corpus of interest (the K in step 6 of figure 3.). Widely used *measures of dependence*⁶ [1] include χ^2 , mutual information, correlation coefficient, cosine,

⁶ sometimes also called *measures of (im)purity* [7] or *measures of interestingness* [5,15,16]

Gini index and many others (see [5,15,16] for several insightful comparisons among them).

In order to experiment with a large number of such co-occurrence measures we have used Christian Borgelt's INES package [1]. This package implements over 20 such measures for which it can construct various graphical models such as optimal spanning trees. Experimenting with such a large number of measures we have observed that:

1. measures tend to produce trees with characteristic shape types, with width vs. depth being a salient dimension for differentiating them;
2. deep and narrow trees tend to be inferior as a source of *connecting terms* because
 - (a) deep trees yield large sets of *connecting terms*
 - (b) large sets of *connecting terms* can start to outweigh the original query terms and lead the IR engine astray; and
 - (c) large singly connected graphs, from which the *connecting terms* are derived, are more likely to contain a low weight link and thus not be coherent units as desired.⁷

Shallow trees tend to be produced by the measures *weight of evidence*, *quadratic information gain*, *relief* and *relevance*, while deep trees tend to be produced by *information gain*, *stochastic complexity* and *reduction of description length*. In addition, we have found that *weight of evidence* [9,7] tends to outperform the other measures in its group because the small set of terms it produces contains terms that one would intuitively recognize as more relevant.

3.3 Selecting Initial Co-occurrence Data

Because it is impractical to compute co-occurrence statistics for all terms in the corpus, our methods includes a step in which a selection of passages is retrieved over which such statistics are computed. This step is the most variable, and in some sense unsatisfactory aspect of our approach. The possible variations are too numerous to discuss in detail here. Here we simply state the choices we settled on after extensive experiments.

The co-occurrence data is selected from a small set of documents that match the original query sufficiently well as is done in pseudo-relevance feedback and local context analysis [19]. Currently we retrieve a fixed number of documents (25), but until recently we were setting the cutoff dynamically based on the

⁷ This is a generalized case of a chain being only as strong as its weakest link.

differences amongst the scores⁸ that the IR engine assigns to the top ranking documents. Further investigations would be needed to gather more empirical evidence on which strategy is best.

The retrieved documents are split into sentences, and a sentence-word matrix is constructed.⁹ Only a top fraction of the sentences, as measured by their *overlap* [8] with the original query, is retained. Using sentences as the basis of co-occurrence is done because of the structure of news corpora, where documents tend to have tens of sentences, with most being irrelevant to the query. Having a finer grain of co-occurrence and discarding the irrelevant sentences from the documents leads to better connecting terms. This process may be inappropriate for a more focused corpus, such as the CNS corpus, which seems to average only about 6 sentences per document, with highly focused documents.

Repetition counts are discarded and only presence/absence information is retained. This is done both because repetition within a single sentence is not significant and because some dependence measures are designed for boolean data. Only terms that occur in a sufficient number of sentences and original query terms are retained for the computation of the connecting terms. This reduces dimensionality, helping prevent overfitting as well as speeding up computation.

3.4 Obtaining final answers

As discussed earlier returning the connecting terms by themselves is typically of limited use to a user. Instead we construct a new query, consisting of the original query terms and the *connecting terms*, with the original terms given greater weight. We retrieve a number of best matching sentences. Since the AQUAINT corpus contains large numbers of nearly identical passages, we have found it necessary to cluster the retrieved sentences using *overlap*, and return only one representative per cluster. This is designed to produce a set of sentences that between them contain all the query terms, without too much redundancy between the sentences.

4. Related Work

While our method bears superficial resemblance to local context analysis [19], it seems conceptually closer to mining for indirect associations [17].

⁸ Kullback-Leibler (or KL) divergence of the retrieved document from the query.

⁹ We effectively treat sentences as documents in their own right.

Local context analysis expands queries with terms that are correlated to one or more query terms. However, the scoring formula does not appear to take the higher order dependency structure into account. In other words, given the query terms *dog* and *cat* local context analysis and pseudo relevance feedback are more likely to select terms that are strongly connected to a part of the query like *poodle*, *bone*, *siamese*, *mouse*, then terms that are moderately connected to all (or most) of the query like *veterinary*, *pet*, *animal*¹⁰ which our method is designed to favor. This bias towards connecting terms is shared by Tan, Kumar and Srivastava [17].¹¹ However, our approach can find linear chains of terms where each term is only correlated to its neighbors. This makes it at least in principle possible to trace a more complex relation between Alice and Emily like

- Alice lives next to Bob
- Bob works with Cathy
- Cathy carpools with Dick
- Dick is Emily's cousin

provided the pairwise correlations exist in the corpus.

It also extends naturally to connecting more than two query terms since we actually work with trees. Tan, Kumar and Srivastava only look for one step chains.

An experiment worth conducting would be to use our algorithm with Tan, Kumar and Srivastava's *IS* similarity measure,¹² which they show to have desirable properties for detecting associations.

There have also been other applications of graphical models to information retrieval [3,18] as for example in the Inquiry system [3]. However, they are used at different levels of abstraction, with nodes representing entire documents and queries, rather than individual term occurrences.

5. Empirical Evaluation

¹⁰ This example is hypothetical, intended to leverage the reader's intuitions about likely strength of associations. It would be an interesting research problem to actually *define* a protocol that would allow to test these intuitions empirically and to use that protocol to test if there really is a significant difference between the technique we describe in this paper and the other techniques to which we are comparing it.

¹¹ We have become aware of this work only recently.

¹² This measure is not among the measures provide by the Bayes net package we currently use. The experiment would thus require us to modify the internals of the INES package.

Our system participated in the relationship track evaluation pilot conducted by the AQUAINT program.¹³ Two assessors independently judged answers to 100 relationship questions, assigning to each answer a score between 0-4, inclusive. Zero was the low score: it meant that the response had no value at all. Four was the top score, and meant that the answer was completely satisfying. The other values had no specific meaning attached. The assessors were explicitly instructed to judge for content, not form, of the response. Apart from the three submissions to the pilot the assessors were also given an answer set supplied by the human subjects that had created the questions. The assessors were unaware of the identities of the runs or even of the presence of a human generated run. The following is a random sample of questions from the exercise:

1. In what country and by whom did Operation Turquoise take place?
2. What is the purpose in taking DHEA?
3. Why is the International Olympic Committee (IOC) concerned about human growth hormone?
4. What part did ITT (International Telephone and Telegraph) and Anaconda Copper play in the Chilean 1970 election?
5. What's the connection between the United States Navy and Puerto Rico's Vieques Island?
6. Who are the leaders of the Lebanese guerrilla organization Hezbollah?
7. What is the relationship between the West Nile Virus and the St. Louis encephalitis?
8. What caused the government to sue Microsoft?
9. What effect did introduction of the "iMac" computer have on the fortunes of Apple Corp.?
10. What is the connection between Jesse Ventura and Target Stores?
11. Why does Romania have a non-Slavic name?
12. What has been the relationship between Osama bin Laden and Sudan?
13. Was London's Millennium Dome successful as a tourist attraction?

In all results presented below our QUIRK system is system A.

Human	System A (QUIRK)	System B	System C
308.5	165.5	48.5	34.5

Figure 4: Total scores averaged over assessors

¹³ We thank Ellen Voorhees, who organized the relationship question evaluation at NIST, for giving us permission to use the evaluation data in this paper.

The total scores for the runs averaged between the two assessors are given in figure 4.

Run	Assessor 1		Assessor 2	
	0	>0	0	>0
human	2	98	11	89
A (QUIRK)	22	78	45	55
B	76	24	79	21
C	78	22	87	13

Figure 5: Zero vs. Nonzero scores by the assessors

Figure 5 gives the zero/non-zero score counts given by the two judges.

Run	Min	Max	Median	Mean
humans	16	1007	151.5	200.0
A (QUIRK)	32	2844	757.0	832.1
B	4	401	48.5	86.8
C	2	863	15.0	33.2

Figure 6: Answer length statistics measured in non-white-space characters

We must stress that these results must be interpreted against the background of an important design decision. While systems B and C tried to restrict their answers to either a single sentence or a short list of words describing the relationship¹⁴ (comparable to our *connecting terms*), we decided, based on the reasoning in section 2.1, to display them in the context of complete sentences, as shown in figure 1. This accounts for the disparity in answer length, and makes it harder to draw conclusions from the scores. The developers of C confirmed our reasoning by performing their own evaluation, comparable to that in figure 5, but also including a significantly higher scoring run of their system with answers presented in context (Run C-J). Their results of are presented in figure 7.

Run	0	>0
A (QUIRK)	39	61
B	80	20
C	78	22
C-J	59	41

¹⁴ Personal communications.

Figure 7: Team C's reevaluation

Run C-J had a median response size of 263 and mean of 561.

6. Conclusion and Further Work

The system, originally tuned to work on the AQUAINT corpus, and tested on the 100 questions in the AQUAINT relationship pilot as discussed above,¹⁵ appears to work well on new corpora and unseen questions without recalibration. We use the same settings for querying the CNS abstracts data but with different questions, appropriate to this data set without apparent loss in performance. We also use the query expansion method as a part in other evaluations, where we have observed it to be beneficial. Thus it appears at least that we have not overfitted the initial question set, which was quite diverse to begin with. This lends some further empirical support to the underlying ideas of our approach. However, many incidental decisions had to be made to turn these ideas into a working system. Most of these were arrived at by trial and error. More data on comparisons with other system is also desirable.

7. Acknowledgements

I would like to thank the other QUIRK members Stefano Bertolo and Bjoern Aldag for their feedback on the ideas leading to this approach and extensive editorial work on this paper. We wish to acknowledge the support of the ARDA AQUAINT program, which has funded this research. We would also like to thank the anonymous reviewers for their constructive suggestions. Much credit is also due to the Open Source software community that has provided many critical components of our system: specific tools like Lemur and INES, as well as the ubiquitous infrastructure of Linux, GNU, Python, Perl ...

References

- [1] C. Borgelt. *Data Mining with Graphical Models*. PhD thesis, Otto-von-Guericke-University of Magdeburg, Germany, 2000.
- [2] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic query expansion using SMART: TREC 3. In *Text REtrieval Conference*,

¹⁵ See Appendix for a sample of the questions in the pilot.

pages 0–, 1994.

- [3] James P. Callan, W. Bruce Croft, and Stephen M. Harding. The INQUERY retrieval system. In *Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications*, pages 78–83, 1992.
- [4] Eugene Charniak. A neat theory of marker passing. *AAAI-86*, 1: 584–588, 1986.
- [5] Robert J. Hilderman and Howard J. Hamilton. Heuristic measures of interestingness. In *Principles of Data Mining and Knowledge Discovery*, pages 232–241, 1999.
- [6] Jaana Kekäläinen and Kalervo Järvelin. The impact of query structure and query expansion on retrieval.
- [7] Igor Kononenko. On biases in estimating multi-valued attributes. In *IJCAI*, pages 1034–1040, 1995.
- [8] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.
- [9] D. Michie. Personal models of rationality. *Journal of Statistical Planning and Inference*, 1988.
- [10] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *Research and Development in Information Retrieval*, pages 206–214, 1998.
- [11] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, revised second edition, 1994.
- [12] Judea Pearl. Graphical models for probabilistic and causal reasoning. In Dov M. Gabbay and Philippe Smets, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Volume 1: Quantified Representation of Uncertainty and Imprecision*, pages 367–389. Kluwer Academic Publishers, Dordrecht, 1998.
- [13] M. R. Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 227–270. MIT Press, Cambridge, 1968.
- [14] R. Sedgewick. *Algorithms*. Addison Wesley, 1983.
- [15] P. Tan and V. Kumar. Interestingness measures for association patterns: A perspective, 2000.
- [16] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns, 2002.
- [17] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Indirect association: Mining higher order dependencies in data. In *Principles of Data Mining and Knowledge Discovery*, pages 632–637, 2000.
- [18] Konstadinos Tzeras and Stephan Hartmann. Automatic indexing based on bayesian inference networks. In Robert Korfhage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of SIGIR-93, 16th ACM International*

Conference on Research and Development in Information Retrieval, pages 22–34, Pittsburgh, US, 1993. ACM Press, New York, US.

- [19] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, 1996.