Automatic thesaurus construction

A paper written within the GSLT course Linguistic Resources, autumn 2002

Monica Lassi
monica.lassi@hb.se
Graduate School of Language Technology
Swedish School of Library and Information Science,
University College of Borås

Abstract
One of the major problems of modern Information Retrieval (IR) systems is the vocabulary problem that concerns the discrepancies between terms used for describing documents and the terms used by the searchers to describe their information need. A way of handling the vocabulary problem is by using a thesaurus, which shows (usually semantic) relationships between terms. Three approaches for automatically creating thesauri are presented in this paper; statistical co-occurrence analyses, the concept space approach, and Bayesian networks.

1. Introduction

A thesaurus (plural: thesauri) is a valuable tool in Information Retrieval (IR), both in the indexing process and in the searching process, used as a controlled vocabulary and as a means for expanding or altering queries. Most thesauri that users encounter are manually constructed by domain experts and/or experts at document description. Manual thesaurus construction is a time-consuming and quite expensive process, and the results are bound to be more or less subjective since the person creating the thesaurus make choices that affect the structure of the thesaurus. There is a need for methods of automatically construct thesauri, which besides from the improvements in time and cost aspects can result in more objective thesauri that are easier to update.

This paper is written within the course Linguistic Resources. The objective of this paper is to present an overview of how thesauri are used in IR and how thesauri can be created automatically. First, some of the aims of the Information Retrieval (IR) field is presented, along with some of the research problems of the field. Second is a presentation of thesauri – what they are and how they are used in IR. Third is an overview of some approaches to automatic thesaurus construction.

2. Information Retrieval

Information Retrieval (IR) research is an empirically driven field of research with its origins in the 1960s. The goal of IR research is to develop systems that retrieve information that is relevant to an information need. Two of the most important features of an IR system are that it retrieves all information relevant to the information need of the user, and that it only retrieves relevant information. Other important factors are the efficiency of storage and searching, i.e. that the information is accurately stored when a minimal amount of disk space is used and the system retrieves information as quickly as possible. The term information is in this case used in quite a broad sense, and denotes the traditional meaning of text stored in documents, as well as images, sound and moving pictures.

There are two main operations in IR; document analysis and matching. The document description operation creates a representation of each document that is to be included in an IR system. The representations facilitate searching and divide the document into different fields that are used in the matching operation. Some fields are used to describe the form of the document (e.g. author, publisher, journal name, and publication year), while others describe the content of the document, i.e. what the document is about. There are different kinds of content descriptions, such as indexing, classification, and clustering. All of these can be performed manually or automatically. The document descriptors are stored in the IR system and matched against the query of the user. An information need must be transformed from the user's natural language to a query that the system can understand and process. Documents that have document descriptors that match the user's query are retrieved and presented, and are usually ranked according to the documents' relevance to the query.

Evaluation of IR systems are usually performed by measuring the efficiency and effectiveness of the system.. There are two major aspects of efficiency: time and space– how fast the system can match queries to document descriptions, and how much disk space the system requires. The data structure used for storage has a big impact on both storage space and the time aspect. The effectiveness aspect determines how well the system retrieves documents that are relevant to a query. The optimal result is that all documents retrieved are relevant to the query, and that no non-relevant documents are retrieved. There are a few problems with this idea, one being that the determination of relevance is a very subjective process that is closely related to factors such as the situation of the user at a specific place and time. A person who is a novice in space physics but wants to learn more, will probably find information on general space physics more relevant than information on e.g. electron spectrometers for nanosatellites. At a later stage, when the person has become more of an expert on the subject, the general information is no longer relevant, and information on electron spectrometers for nanosatellites may very well be more relevant. A document collection used for evaluation of IR systems has an assessment of each document's relevance to a specific topic. Taken into consideration that humans perform the assessments, the collection may include some assessments that are not correct – inconsistency and subjectivity may possibly affect the outcome of the evaluation. This problem might be particularly apparent in a paradigm that, at least to its researchers, is considered objective.

The most commonly used measurements of retrieval performance are precision and recall. Precision measures the ability of the system to retrieve only the documents that are relevant to a query:

$$\text{precision} = \frac{\text{amount of relevant documents retrieved}}{\text{amount of documents retrieved}}$$

Recall measures the ability of the system to retrieve all documents that are relevant to a query:

$$\text{recall} = \frac{\text{amount of relevant documents retrieved}}{\text{amount of relevant documents in the collection}}$$

2.1. Indexing

The indexing process is part of the document description process, and concerns content analysis and description. This process can be performed by humans – manual indexing, or by computer software – automatic indexing. The aim of indexing is to assign such words to each document in a collection that the contents of the document are sufficiently disclosed by these words. The assigned words are called index terms since they are stored in the index and used in the matching process. Most of the commercial databases that sell scientific information (e.g. LISA, Inspec, ERIC, etc.) are based on manual indexing. Some of these databases might use automatic indexing as a complement, to enable users to search entire documents, regardless of the index terms used to describe the documents. In contrast, the indexing carried out by Web search engines is carried out with relatively crude word counting techniques, with no human involvement. It is not possible for humans to process each document published on the Web, which leaves the methods used by Web search engines as the only alternative. Regardless of the indexing being performed manually or automatically, the index terms are linguistic entities, which are to be used by searchers. Many IR systems require exact queries to retrieve relevant information, and the user must therefore have a fairly good idea of which index terms that are available and, if possible, how the terms are related.

The most commonly used methods in automatic indexing are based on weighting words according to their frequency of occurrence in documents. The better a term is considered to be at representing the content of a document, the higher weight it will be assigned. Most IR researchers seem agree on the importance of using a good term weighting scheme, since the weights of terms and documents can be matched to the weight of each term in a query, as well as for ranking retrieved documents according to their relevance ranking. Quite a large amount of effort has gone into refining different weighting formulas, and the refinements have increased performance to a certain extent (Anderson & Pérez-Carballo, 2001, p. 260). Some of the weighting formulas do not have much theory to support them, and are rather the results of trial and error (ibid).

Term Frequency (TF), Inverse Document Frequency (IDF) and a combination of the two, TF*IDF, are the most commonly used weighting techniques:

- Using TF, the optimal index terms are those that occur with a medium frequency in a document. The most frequent words are considered bad discriminators – they cannot discriminate one document from the rest of the collection. The least frequent words are considered too insignificant to be good content descriptors.
- While the TF technique considers each document by itself, the IDF technique takes the word occurrences in the entire collection into account. A word that occurs in all or many documents is not good at discriminating documents from each other. On the other hand, a word that occurs in few documents in the collection are considered good index terms, since the word may clearly discriminate a few documents from the rest of the collection.
- When TF and IDF are combined into TF*IDF, both the occurrence of words in each document, and the occurrence of words in the document collection are taken into account. The result is high weights for words that occur with medium frequency in an individual document and with low frequency in the collection.

The methods described above are purely statistical, and do not consider the linguistic properties of the documents. For instance, basic TF*IDF concerns single word index terms, but support for phrasal index terms is usually implemented. Phrases are not alike, and features such as position within the text and type of phrase affect the derivation of phrasal index terms (Strzalkowski 1995, p. 401). Strzalkowski claims that compound terms

derived by means of word counting behave differently from terms derived by means of e.g. part-of-speech analysis. He continues by claiming that "*…the lack of an established retrieval model that can handle linguistically motivated compound terms may be most serious obstacle in evaluating the impact and feasibility of NLP in IR*". (ibid)

2.2. Vocabularies and terms

In IR, the distinction between controlled and uncontrolled vocabularies is often made. Uncontrolled vocabularies allows for every token in a document to be a potential index term, without paying more than minimal attention (or none at all) to word form and other linguistic features. Controlled vocabularies on the other hand, have rules that regulate which words that are allowed to be index terms, as well as the word forms and other specific features of those terms. Thesauri are a great help for indexers when determining index terms for documents, showing which terms that are related in some way, with references from a disallowed term to the preferred term etc. Searchers can use controlled vocabularies to determine how to translate their information need to the language of the database. When automatic indexing is used, the vocabulary is often uncontrolled. There is, however, a movement towards using thesauri in automatic indexing as well, which has proven to be especially useful when the documents indexed are in the same domain.

Miller (1997, p. 482) lists some arguments for using a controlled vocabulary:

- Free-text search (text indexed with little or no lexical control) is only useful when seeking information on "super-new or super-narrow problems" which have not yet been incorporated into the controlled vocabulary.
- Free-text search leads to errors due to rich variations in the language, such as synonymy, different word forms, spelling variations etc.
- Whether a searcher is capable at searching in a database depends on how much instructions are given. A controlled vocabulary usually have well-defined rules for how terms are created, making it easier for the searcher to find the correct terms.

Controlled vocabularies can be either pre-coordinated or post-coordinated. In pre-coordinated systems, index terms are coordinated by the indexer (or by the IR system when automatic indexing is used) at indexing time. In post-coordinated systems, index terms are coordinated at search time – the searcher combines the index terms that make up the query.

3. Thesauri

A thesaurus is a controlled vocabulary that shows relations (e.g. semantic) between terms, which can aid searchers in finding related terms to expand queries. There are many different definitions of thesauri, varying from quite modest definitions that focus on the relations between words without stating which kinds of relations that are meant, to such definitions that state more exactly which relations that are concerned. An example of quite a modest definition is presented by Schütze and Pedersen: "*We define a thesaurus as simply a mapping from words to other closely related words*". They continue by stating that a thesaurus must be specific enough to present the searcher synonyms of words in the corpus that is searched. (Schütze & Pedersen 1997, p. 307) Moreover, they state that a thesaurus must cover all of the words found in queries (ibid), which is an interesting statement – how can we know in advance which words are to be used for searching?

In contrast to Schütze and Pedersen, Miller gives a more elaborate definition of a thesaurus as "a *lexico-semantical model of a conceptual reality or its constituent, which is expressed in the form of a system of terms and their relations, offers access via multiple aspects and is used as a processing and searching tool of an information retrieval unit. Hence, it appears that a process of thesaurus construction is a process of simulation in a lexical form: of the whole universum of realities and concepts or its part of hierarchical and associative connections and relations between these realities and concepts*." (Miller 1997, p. 489) Miller states that many professionals working with document description and in related areas, have misunderstood the very essence of the thesaurus when they make a distinction between the thesaurus as a theoretical model and as a practical tool for IR. He means that the two roles are inseparable and that such distinctions are a reflection of "psychological problems connected with the whole complex of information service and expressed sometimes in blind idolization of a user." (ibid)

The relationships between one term and another are often denoted as narrower term (NT), broader term (BT), preferred term (USE), etc. The RT relationship (related term) denotes relationships that cannot be described as a

clearly narrower or broader semantic relationship, which can lead to strange relations. An example of a problem with using the RT relationship is described by Harter and Cheng (1996, p. 312). In the Thesaurus of ERIC (The Educational Resources Information Center) Descriptors, the terms 'literature' and 'humanism' have a RT relationship with each other. To Harter and Cheng this relationship is quite weak and far fetched in a database such as ERIC: "/…/ *most information seekers interested in the idea of humanism are unlikely to find the descriptor LITERATURE especially helpful in building a successful search strategy*." (ibid)

Computational linguistics

Broader Terms
Linguistics

Narrower Terms
Machine Translation

Related Terms
Automatic Indexing
Computer Science
Language Processing
Linguistic Theory
Mathematical Linguistics
Mathematical Logic
Programming Languages
Semantics
Statistics
Structural Analysis (Linguistics)
Word Frequency

Figure 1: Example of the term 'computational linguistics' in the Thesaurus of ERIC Descriptors

Thesauri have been an important tool in modern IR for a long time. Early use of thesauri in IR focused on aiding indexers in the process of doing consistent document descriptions by providing them with accepted index terms. Later on, thesauri have become an aid in the retrieval process as well. Shiri and Revie suggest that the refocusing of the use of thesauri means that it is clear that professionals see the potential of using thesauri in one of the largest IR environments, i.e. the Web. (Shiri & Revie 2000, p. 273)

The vocabulary problem is the notion of people using different words for describing the same concept. An experiment by Furnas, Landauer, Gomes and Dumais[i] showed that when a group of people were to spontaneously choose words to describe objects in five domains, the probability for two people choosing the same term was less than 20% (Chen et al. 1997, p. 17). Chen et al. describe the indexing and search uncertainty in information science as the primary source or IR problems (ibid). It is commonly known in the IR community that indexers to a large extent assign different terms to the same concept, and that the same will happen when an indexer indexes the same document at different times. Due to different backgrounds, experiences etc., searchers will also use different terms to describe the same information need, and all these discrepancies result in a decrease in retrieval effectiveness.

Chen et al. (1995, p. 177) claim that the vocabulary problem is particularly challenging in a scientific community due to the diversity of specialization within domains. This is particularly problematic in such domains where scientific discoveries create needs for new terms to be added to the vocabulary. A related question that Chen et al. pose is how searchers who are not familiar with a specific subject area and its terminology will be able to express their information need (ibid). They believe that their concept space model, which is described later in this paper, is the solution to this problem (ibid, p. 191).

The conventional way of constructing a thesaurus is to do so manually, by building a semantic mapping table. This method requires an expert of the domain that the thesaurus is constructed for, or an expert of document description and IR. Since it is a quite time-consuming process, it is also expensive, and it is hard to know if the

use of the thesaurus will justify the cost of the construction. One way of cutting costs is to reuse an existing lexicographic database, such as WordNet. The problem with this approach is that these databases are general, whereas domain-specific databases need a more specified vocabulary (Schütze & Pedersen 1997, p. 308).

3.1. Approaches to automatic thesaurus construction

3.1.1. Co-occurrence analysis

Term co-occurrence analysis is one of the approaches used in IR research for forming multi-phrase terms. Another approach is part-of-speech analysis, which takes the linguistic properties of the texts into account. Co-occurrence analysis, on the other hand, is a statistical approach where the occurrences of terms in documents, chapters or some other unit are computed. The closer the words occur, the more significant is the co-occurrence. Many automatic indexing methods do not consider how closely words occur, just if they occur in the same document.

The method used by Schütze and Pedersen starts with creation of a matrix that displays the number of times a word co-occurs with other words in a document, in a chapter or in a window of a number of words. Schütze & Pedersen describes this matrix as a "*term-by-term matrix* C *where element* Cij *records the number of times that words* i *and* j *cooccur in a window of size* k */.../. Topical or semantic similarity between two words can then be defined as the cosine between the corresponding columns of* C. *The assumption is that words with similar meanings will occur with similar neighbours if enough text material is available*." (Schütze & Pedersen 1997, p. 311) There are efficiency problems with this approach: the matrix that is used to compare each word in the vocabulary to all other words in the vocabulary tend to be quite large, and it takes quite a long time to process the word comparisons, depending on the size of the vocabulary.

Lexical co-occurrence analysis is closely related to Latent Semantic Indexing (LSI). In LSI, document-by-word matrices are created and processed by Singular Value Composition (SVD) instead of word-by-word matrices. A system based on word-matching will not be able to retrieve documents on 'cosmonauts' in response to a query about 'astronauts'. This problem is though handled with LSI. Schütze and Pedersen explain that there are two major differences between lexical co-occurrence and LSI. The first is of technical nature, and has to do with time complexity. LSI analysis is more time consuming than SVD, the latter not having as strong relation between time complexity and corpus size as the former. The second difference is of conceptual nature and concerns the fact that the lexical co-occurrence approach uses term representations independently, whereas in LSI term representations are only used to compute document representations. The lexical co-occurrence approach thus exploits more of the information in the term vectors than LSI does. (Schütze & Pedersen 1997, p. 310f)

Chen et al. (1995, p. 178) declare that the research in the area of generating terms by co-occurrence analysis done in the 1980's and 1990's have not given the results hoped for. Some experiments have resulted in poor retrieval when terms were generated completely automatically. An experiment by Ekmekcioglu, Robertson & Willet[ii] in 1992 employed a user-directed approach where users were suggested terms for expanding their queries. Four different approaches were used; original queries; query expansions generated by co-occurrence data; Soundex codes, where the same phonetic code is assigned to words that sound the same; and strong similarity measure based on similar character microstructure. The results showed no significant difference in retrieval effectiveness between the initial queries and the expansions made. There was, however, a very small degree of overlap between the relevant documents retrieved by the initial queries and the relevant documents retrieved by the co-occurrence approach. The small degree of overlap of retrieved documents show that by adding to the query terms from a thesaurus generated by co-occurrence analysis, an increase in the number of relevant documents retrieved can be accomplished. (ibid, p. 178f)

3.1.2. The concept space approach

The concept space approach for automatic thesaurus generation was developed by Chen, Ng, Martinez, and Schatz.. A *concept space* is defined as a network of terms and weighted associations which can represent concepts (terms) and their associations for the underlying *information space* which represent the documents in the database. (Where nothing else is stated, the source of this section (3.2.2.1) is Chen et al 1997, p. 21-23)

The concept space approach consist of four steps: (1) Document and object list collection; (2) object filtering and automatic indexing; (3) co-occurrence analysis; and (4) associative retrieval. These steps are presented below.

Document and object list collection

The first effort in creating a thesaurus is to identify the document collection to build the thesaurus on. As Chen et al. wished to create a domain-specific thesaurus, it was important for the collection to consist of a complete and recent collection of document in the specific domain, which could serve as a representative source for the vocabulary to be created. Domain-specific keywords in databases can be used to automatically identify important concepts in documents, and Chen et al. used four different sources in the worm research area as a basis, collecting lists of researcher names, gene names, experimental methods, and subject descriptors.

Object filtering and automatic indexing

In the object filtering process, terms that matched the terms in the vocabulary created in the former stage were identified. Since some concepts may be missing in the object list collection, an automatic indexing process was also carried out. The indexing method used was one suggested by Salton[iii] and included the following steps: (1) dictionary lookup to identify individual words; (2) filtering of words through a stop-word list to remove words that do not qualify as indexing terms; (3) word stemming to identify word stems for the remaining words; and (4) term-phrase formation to create phrases by combining adjacent words. In this case, the weighting process is not a part of the indexing process, and is found in the following step, the co-occurrence analysis.

Co-occurrence analysis

The co-occurrence analysis started with computations of each term's document frequency (the number of documents in a collection in which a word occurs) and term frequency (the frequency of occurrence of a word in a document). Terms appearing in the title of a document were assigned higher weights than terms in the abstract or other parts of the document. Terms that had been identified by the object filters in the first step were also assigned higher weights than those identified in the automatic indexing process. The inverse document frequency was then computed with some extra features. Multiple-word terms were assigned higher weights than single-word terms since the former usually convey more precise semantic meaning than the latter. Next was the co-occurrence analysis based on the asymmetric "Cluster Function" developed by Chen & Lynch[iv], which in the same paper was proven more effective for representing term association than the frequently used cosine measure.

General terms that appeared in many places in the co-occurrence analysis were penalized, and weighting schemes similar to the weighting method inverse document frequency was used. After consulting experts in the domain of the thesaurus, a maximum number of 100 links for a related term was determined. This number removed approximately 60% of the less relevant co-occurrence pairs.

Associative retrieval

The notion of Anomalous States of Knowledge (ASKs) model is that users of IR systems have an information need, which they present to the system as a problem statement. Inherent in these information needs are ASKs, and the searcher's state of knowledge can be represented as a network of associations between words. The structure and characteristics of the network can help to identify anomalies in the state of knowledge. According to Chen et al., the model contributed to associative indexing and term-association based retrieval. Many models of human knowledge and memory associations are represented by network-like structures. Anderson[v],[vi] proposes that people remember the meaning underlying a passed verbal communication, and not the exact wording of it. The proposition is the smallest unit of knowledge that can stand as an assertion bearing meaning. The memory can be thought of as a network of such propositions. The strength of the association paths that lead to a piece of information contributes to the level of activation being spread. This theory, that Chen et al. call *spreading activation* is the influence of many semantic network-based IR systems. (Chen et al 1997, p. 19)

A browsing interface was created for the thesaurus, as well as possibilities for users to invoke selected spreading activation algorithms for multiple-term, multiple-link term suggestions. One of the algorithms has proven especially effective for concept-based IR, i.e. the Hopfield net algorithm, which is a neural network that can be used as a content-accessible memory. "*Knowledge and information can be stored in single-layered, inter-connected neurons (nodes) and weighted synapses (links) and can be retrieved based on the Hopfield network's parallel relaxation and convergence methods. The Hopfield net has been used successfully in such applications as image classification, character recognition, and robotics and was first adopted for concept*-based IR *in Chen et al. 1993*[vii]."

Each term in the network-like thesaurus was treated as a neuron and the asymmetric weight between any two terms was taken as the unidirectional, weighted connection between neurons. Using user-supplied terms as input patterns, the Hopfield algorithm activated their neighbours (i.e. strongly associated terms), combined weights

from all associated neighbours and repeated this process until convergence. During this process, the algorithm caused a *damping effect,* in which terms farther away from the initial terms received gradually decreasing activation weights and activation eventually "died out". This phenomenon is consistent with the human memory *spreading activation* process.

3.1.3. Bayesian networks

The major problem of statistical methods for automatic thesauri construction is according to Park and Choi (1996, p. 544) data sparseness. Despite many attempts to solve this problem, no successful method has been found. They have experimented with Bayesian networks in an attempt to find a way of handling the problem, and explain the approach: "*A Bayesian network built from local term dependencies can give a probabilistic similarity distribution among the terms. The distribution is different from that of the most frequent probabilities, but will serve the purpose of classifying terms.*" (ibid)

A short introduction to the basics of Bayesian networks will be given below, to make it clearer how Bayesian networks can be used to analyze co-occurrences and to generate a thesaurus with terms with high co-occurrence. Charniak (1991) introduces Bayesian networks for beginners by presenting a directed acyclic graph (DAG) that represents a situation in which causality plays a part, but where the actual order and occurrences of events are not certain and the situation has to be described probabilistically. The situation involves a man who goes home at night, and wants to know if his family is home before he tries to open the door. When his wife leaves the house, she turns on an outdoor light. She also does this when she is expecting a guest or when the dog is outside and nobody is at home or when the dog has bowel troubles. If the dog is outside, it usually barks when someone comes, but sometimes it is hard to know if it is the family's dog or some other dog in the neighbourhood. Sometimes someone has left the house without turning on the light or letting the dog out. In cases like this, it is hard to know what to infer, not all evidence points in the same way. (Charniak 1992, p. 51) This situation can be represented by this graph:

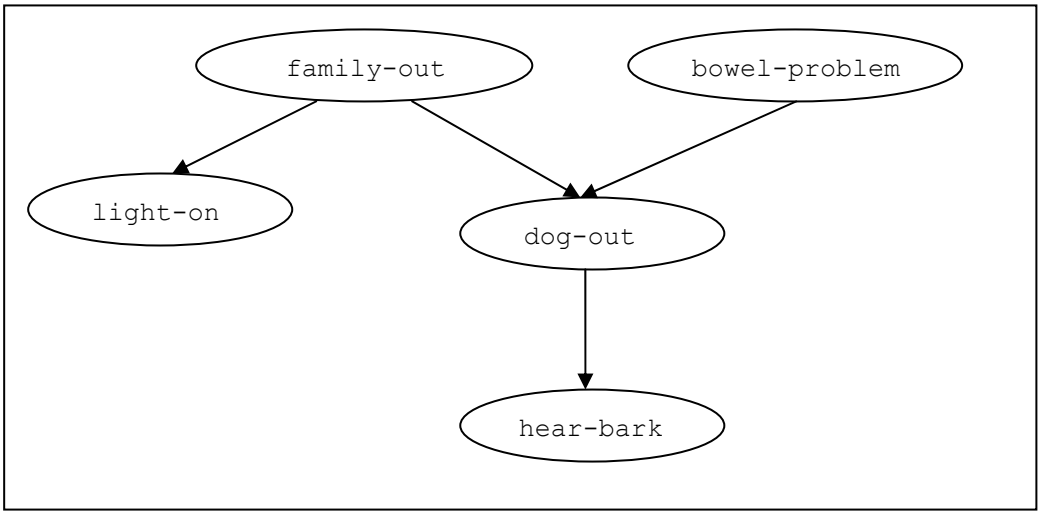

Figure 2: A directed acyclic graph representing the states described in the text (Charniak 1992, p. 51)

The nodes of the DAGs are random variables that often represent states and can take values such as true/false, discrete as well as real numbers etc. The arcs specify the independence assumptions between the random variables. The assumptions determine what probability information is required to specify the probability distribution among the random variables. To specify the probability distribution of a Bayesian network, all root nodes must be given prior probabilities (since there is no predecessor to a root node). The next step is to give all nonroot nodes the conditional probabilities given all possible combinations of their direct predecessors. (Charniak 1992, p. 51) The network in the prior image has the following probabilities (ibid, p. 52) :

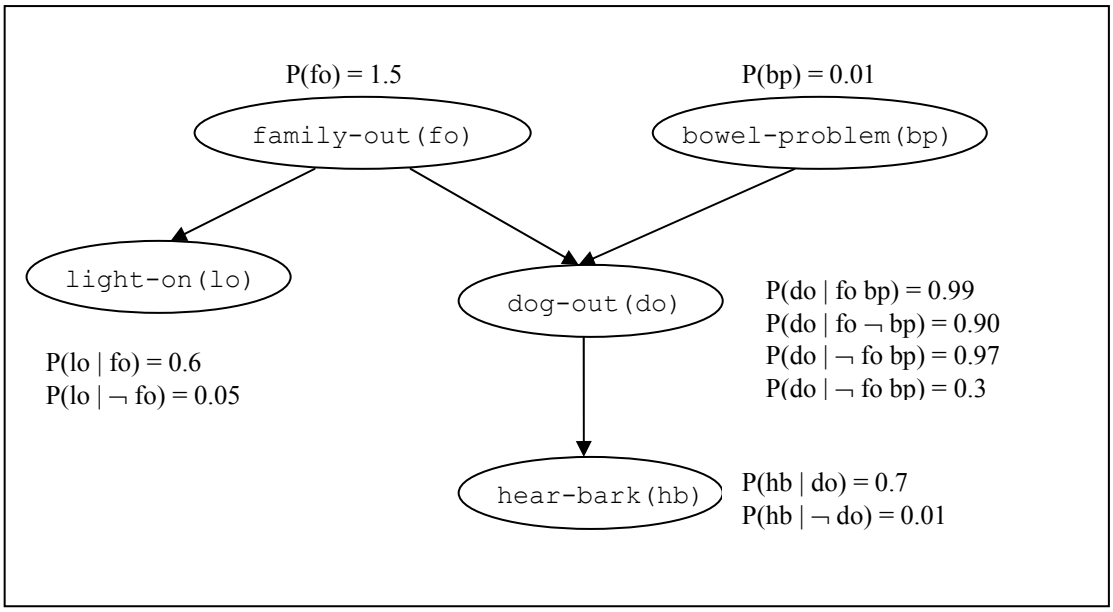

Figure 3: The graph in figure 2 with the probabilities of each state occurring stated (Charniak 1992, p. 52)

The Bayesian network shown above states that family members turn on the light when leaving the house 60% of the time, and that the light will be switched on 5% of the time even though nobody leaves the house. If the dog is outside, one can hear it bark 70% of the time, and it will be quiet only 1% of the time. (ibid)

4. Discussion

One of the major problems of the modern IR systems is the vocabulary problem that concerns the discrepancies between terms used for describing documents and the terms used by the searchers to describe an information need. A way of handling the vocabulary problem is by using a thesaurus, which shows (usually semantic) relationships between terms. Thesauri can aid the indexer or the indexing system in choosing the correct terms to describe the contents of documents, and in normalizing the terms so that all terms are e.g. presented in singular form. In the searching process, thesauri can help the searcher to find terms to refine a query, e.g. by expansion of the original query.

Some of the relationships between terms that are handled by thesauri are narrower term (NT), broader term (BT), preferred term (USE) and related term (RT). There are some obvious problems with manually constructing thesauri. It is an expensive and time-consuming process that requires a domain-expert or an expert at document description. In domains where new research fields develop frequently, thesauri become out of date, and need to be updated, which again is time-consuming and expensive. By using documents published in the domain in question as a corpus, a thesaurus can be created and updated automatically. The terminology of the researchers of the field will be the basis of the indexing process and the assignment of index terms. There are a number of different approaches available for automatically creating thesauri, among others different kinds of statistical co-occurrence analyses, the concept space approach and by representing the terms as Bayesian networks. These quite different approaches have been briefly presented in this paper. A way of following up this paper would be to go in deeper on the different approaches, and/or select the one most interesting for my future thesis project.

5. References

Anderson, J. D. & Pérez-Carballo, J. (2001). "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part 2: Machine indexing, and the allocation of human versus machine effort." Information Processing and Management **37**: 231-254.

Charniak, E. (1991). Bayesian networks without tears. AI Magazine Winter 1991, 50-63.

Chen, H., T. D. Ng, et al. (1997). "A Concept Space Approach to addressing the vocabulary problem in scientific Information Retrieval: An experiment on the Worm Community System." Journal of the American Society for Information Science **48**(1): 17-31.

Chen, H., T. Yim, et al. (1995). "Automatic thesaurus generation for an electronic community system." Journal of the American Society for Information Science **46**(3): 175-193.

Harter, S. P. and Y.-R. Cheng (1996). "Colinked descriptors: Improving vocabulary selection for end-user searching." Journal of the American Society for Information Science **47**(4): 311-325.

Miller, U. (1997). "Thesaurus construction : problems and their roots." Information Processing and Management **33**(4): 481-493.

Park, Y. C. and K.-S. Choi (1996). "Automatic thesaurus construction using Bayesian networks." Information Processing and Management **32**(5): 543-553.

Shiri Asghar, A. and C. Revie (2000). "Thesauri on the Web: Current developments and trends." Online Information Review **24**(4): 273-279.

Schütze, H. and J. O. Pedersen (1997). "A cooccurrence-based thesaurus and two applications to Information Retrieval." Information Processing and Management **33**(3): 307-318.

Strzalkowski, T. (1995). "Natural language information retrieval." Information Processing and Management **31**(3): 397-417.

6. Notes

[i] Furnas, G.W., Landauer, T.K., Gomez, L.M. & Dumais, S.T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*. 30(11), 964-971.

[ii] Ekmekcioglu, F.C., Robertson, A.M. & Willett, P. (1992). Effectiveness of query expansion in ranked-output document retrieval systems. *Journal of Information Science*. 18, 139-147.

[iii] Salton, G. (1989). *Automatic text processing*. Reading, MA: Addison-Wesley.

[iv] Chen, H & Lynch, K.J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transacations on Systems, Man and Cybernetics*. 22(5), 885-902.

[v] Anderson, J.R. (1985). *Cognitive psychology and its implications (2nd ed.)*. New york: W.H. Freeman and Company.

[vi] Anderson, J.R. (1985). Indexing systems: Extensions of the mind's organizing power. *Information and Behaviour*. 1, 287-323.

[vii] Chen H., Lynch, K.J., Basu, K., & Ng, D.T. (1993). Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Expert. Special Series on Artificial Intelligence in Text-based Information Systems.* 8(2), 885-902.