REASONING ABOUT KEYWORDS USING DEFAULT LOGIC
An Extended Abstract

Anthony Hunter,
Department of Computing, Imperial College, London

12 September 1995

1 Introduction

The aim of information retrieval is to provide a user with the "best possible" information from a database. The problem of information retrieval is determining what constitutes the best possible information for a given user. A common form of interaction for information retrieval is for the user to offer a set of keywords. These are then used by the information retrieval system to identify information that meets the users needs. For example, in a bibliographic database, a user might be interested in finding papers on some topic. The keywords would be an attempt to delineate that topic, and so are used to improve precision (ensuring that a significiant proportion of the items retrieved are relevant to the user) and recall (ensuring that a significant proportion of the relevant items are retrieved). To support this, we need to reason about keywords, to identify for example, meanings, synonyms, related terms, to resolve ambiguities, and more generally, to make the process more knowledge-based. Here, we consider the need for using default logic for reasoning with syntactic, semantic and statistical information about keywords.

2 Statistical analysis versus semantic analysis

In order to determine how well matched an item in a database is for a query, most formal approaches to modelling the uncertainty in information retrieval use statistical information about keywords in the database. For example, a keyword in common between an item and a query, that occurs more in the item than in any other item, is a distinguishing feature of that item, and hence can increase the posterior probability that the item is of relevance to the user. A variety of such discriminating factors based on statistics have been proposed to quantify the similarity between items and queries. A great advantage of using statistical analysis is the volume of data that can be used to identify statistical relationships. For example, consider the amount of data in a bibliographic database for a university library or in a text database of past editions in newspaper editorial office.

In statistical analysis, the relationship between keyphrases is established by frequency ratios, whereas in semantic analysis, the relationship is established by "meaning". Not using semantic information is wasting valuable information that could be critical in matching a users needs to the information in the database.

For example, suppose a search is undertaken using the keyword *car*, it would miss all items that only have the keywords *automobile* and *motor.car*. Similarly, suppose a search is undertaken using the request *computer.network or academic.communications*, an excessive number of items might be retrieved, whereas what might have been actually of interest to the user was the more specialized subject *internet*. Therefore, what is required is some formal representation of the semantic interrelationships between concepts, together with some ability to interpret users intended meanings when presenting requests.

3 The need for more than a thesaurus

Thesauri are widely-used tools in information retrieval for representing semantic information about keywords. Essentially, a thesaurus is a database where for each keyword, there is a listing of synonyms, more specialized keywords, more general keywords, and related keywords. Usually they are not automated tools. Rather they are used directly by the user for consultation, with the onus being on the user to interepret and utilize the information in the course of composing a request. Current thesauri also lack an adequate representation of context dependency.

In order to capture semantic information more fully in information retrieval, we need to automate the information in the thesaurus. In addition, we need more sophisticated information. In particular, we need context sensitivity. This means we need to be able to specify when any particular specialization, generalization, synonym, or related term for a keyword can be used. Furthermore, we need to be able to extend this to resolving ambiguity such as arising from polysemes. For example, suppose we have the keyword *car*, then usually we would usually be interested in the synonym *automobile*. An exception would be if we also had the keyword *railway*. In which case we would usually be interested in the synonym *wagon*.

4 Formalisms for semantic information

There have been attempts to use knowledge-based systems to provide semantic analysis of keywords in information retrieval ([1], [2]). For example, stemming is widely acknowledged as improving performance, but it can introduce ambiguities. Using a knowledge-base about word meanings can minimize these ambiguities being introduced ([3]). As another example, when sufficient contextual information is available in a request, disambiguation of keywords using a semantic network can be important in improving performance, and is likely to be even more important for diverse subject matter ([4]).

Semantics has also been used to enhance statistical techniques. For example thematic roles (from linguistic analysis) have been used to semantically categorize words in a query. This together with an extension of a probabilistic similarity function between queries and items ([5]) has provided a richer, finer-grained, statistical analysis.

There is now a need for a more general framework for representing and reasoning with semantic information in information retrieval. Previous approaches to semantic analysis in information retrieval are either ad hoc representation schemes, or insufficiently general in their approach, or lack some desirable properties. However, it is possible to use (non-monotonic) logics to handle semantic information about keywords, and so to identify logical relationships between items and queries.

5 Default logic for reasoning about keywords

Default logic, proposed by Reiter [6], is a well-understood formalism for representing uncertain information. For a good introduction, and review of the literature on default logic see Besnard, [7]. A number of variants of default logic have been proposed that provide a range of capabilities (for example [8], [9]) and relationships with probability theory are being established (for example Cussens, [10]).

It seems that default logic and its variants offer a formal framework for representing semantic information that is sufficiently general, and yet lucid ([11]). Furthermore, syntactic and statistical information about words could also be represented in default logic, hence giving a uniform and comprehensive framework for reasoning with keywords.

Supporting technology for default logic is now being developed. There are prototype implementations of inference engines for default logic that can be used for developing default logic knowledge-bases ([12], [13],[14]). Furthermore, there are methods for compiling default logic knowledge-bases for viable implementations of developed knowledge-bases.

References

[1] B Croft. Knowledge-based and statistical approaches to text retrieval. *IEEE Expert*, 8–12, April 1993.

[2] B Vickery and A Vickery. Online search interface design. *The Journal of Documentation*, 49:103–187, 1993.

[3] R Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 191–202. ACM Press, 1993.

[4] E Voorchees. Using wordnet to disabiguate word senses for text retrieval. In *Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 171–180. ACM Press.

[5] E Wendlandt and J Driscoll. Incorporating a semantic analysis into a document retrieval strategy. In *Proceedings of the 14th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'91)*, pages 270–278. ACM Press, 1991.

[6] R Reiter. Default logic. *Artificial Intelligence,*, 13:81–132.

[7] Ph Besnard. *An introduction to default logic.* Springer, 1989.

[8] G Brewka. *Common-sense Reasoning.* Cambridge University Press, 1991.

[9] G Brewka. Reasoning about priorities in default logic. In *Proceedings of the Twelth National Conference on Artificial Intelligence (AAAI'94).* MIT Press, 1994.

[10] J Cussens, A Hunter, and A Srinivasan. Generating explicit orderings for non-monotonic logics. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI'93)*, pages 420–425. MIT Press, 1993.

[11] A Hunter. Using default logic in information retrieval. *Symbolic and Quantitative Approaches to Reasoning and Uncertainty, Lecture Notes in Computer Science*, 946:235–242, 1995. Springer.

[12] I Niemela. A decision method for nonmonotonic reasoning based on autoepistemic reasoning. In *Proceedings of the Fourth International Conference Principles of Knowledge Representation and Reasoning,*, pages 473–484. Morgan Kaufmann, 1994.

[13] T Linke and T Schaub. Lemma handling in default logic theorem provers. *Symbolic and Qualitative Approaches to Reasoning and Uncertainty, Lecture Notes in Computer Science, volume 946*, pages 285–292, 1995. Springer.

[14] T. Schaub. A new methodology for query-answering in default logics via structure-oriented theorem proving. *Journal of Automated Reasoning*. forthcoming.