Knowledge Discovery for Flexible Querying

Henrik L. Larsen, Troels Andreasen, and Henning Christiansen

Department of Computer Science
Roskilde University, DK-4000 Roskilde
{hll, troels, henning}@ruc.dk

Abstract. We present an approach to flexible querying by exploiting similarity knowledge hidden in the information base. The knowledge represents associations between the terms used in descriptions of objects. Central to our approach is a method for mining the database for similarity knowledge, representing this knowledge in a fuzzy relation, and utilizing it in softening of the query. The approach has been implemented, and an experiment has been carried out on a real-world bibliographic database. The experiments demonstrated that without much sophistication in the system, we can automatically to derive domain knowledge that corresponds to human intuition, and utilize this knowledge to obtain a considerable increase in the quality of the search system.

1 Introduction

An increasing number of users access an increasing amount of information with a decreasing insight knowledge about the domains of the information accessed. Indeed, the need of effective solutions to flexible, intutitive querying-answering systems is already obvious, and it becomes increasingly important to provide such solutions, allowing everyday's user find their way trough a mess of information available, in particular through the Internet. For a representative collection of approaches to flexible query-answering we refer to [1–3].

Fuzzy evaluation of queries introduces a flexibility that, as compared to conventional evaluation, brings query-answering a great deal closer to the ideal of human dialog. This flexibility is partly manifested in fuzzy notions available for query specification, and partly in the inclusion and ranking of similar, but non-matching objects in the answer. Since both the interpretation of fuzzy query concepts, and the determination of similar objects, draw on the ability to compare and order values, fuzzy query evaluation is heavily dependent on measures of similarity and on ordering defined for attribute domains underlying the database.

Furthermore, for the flexibility to be apparent from the users point of view, the similarity measure applied must correspond to the user's perception of the attribute domain. That is, two values in an attribute domain are similar to the degree that replacing the one value for the other in an object does not change the user's potential interest in the object. For instance, the two prices of a home, 110,000$ and 115,000$ are rather similar, as are the meanings of the terms 'summer cottage' and 'weekend cottage'.

In numerical domains the similarity measure may often be specified as a function of the distance. For non-numerical domains, such as the domain of subject terms of documents, the similarity measure may in general be represented by a fuzzy relation that is reflexive and asymmetric. Such a relation is depicted by a directed fuzzy graph; when the vertices represent terms, and the edges the (directed) similarity between terms, we refer to such a graph as a *fuzzy term net*.

A central issue in knowledge engineering for flexible querying systems, is the acquisition of the similarity knowledge. Sources for such may include domain experts, and existing ontologies (including thesauri and lexicons). However, other important sources are the information bases queried, and observed user behavior, possibly including feedback from users. To exploit the latter sources, the similarity knowledge must be discovered (or learned) from the set of instances considered. Although similarity knowledge obtained by discovery tends to be less precise, due to its emperical basis, than knowledge aquired from domain experts and ontologies, it has the advantage that it may be applied dynamically to reflect the current use as represented in the information base and by the users of the querying system.

In this paper, we present an approach to discovering and utilizing knowledge in information bases. Especially the discovery aspect, which is also touched upon in [4, 5], is important, since flexible approaches depends on available, useful domain knowledge, and we actually can derive such knowledge from the information base. Of major interest in our approach are domains of words used in describing properties of objects, thus domains with at least no inherent useful ordering. In a library database these may be authors' keywords, librarians' keywords, and words appearing in titles, notes, abstracts or other textual attributes attached to objects.

2 Fuzzy term nets in representation of similarity

We distinguish between two kinds of similarity on domains of words (or terms): thesaurus-like and association-like similarity. A *thesaurus-like* similarity notion relates words with respect to meaning; similar words are words with overlapping meaning, as exemplified in Figure 1(a). The structure depicts a synonym relation, or, more precisely, 'a kind of' relation, on the domain; in the example, for instance that a house is a kind of home. An *association-like* similarity notion relates words with respect to use or ideally with respect to natural human association, as exemplified in Figure 1(b), where a given word points to other words that are similar with respect to the context in which they are applied. We refer to such structures as term nets. Obviously we obtain a generalization when allowing term relationships to be graded, that is, when allowing for fuzzy term nets. A fuzzy term net representing an association structure is referred to as an *associative term net* . The fuzzy relation represented by such a net, reflects the vagueness of the term relationships, and the associated uncertainty due to its emperical basis. For any pair of terms (A, B), the derived strength of a relationship from A to B, $s(A, B)$, is taken as an estimate of the degree to which A is associated

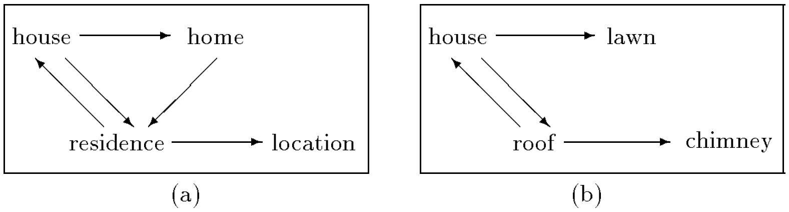

Fig. 1. Term similarity structures: (a) a thesaurus structure, (b) an association structure

to B, such that $s(A, B) > s(A, C)$ indicates that A is associated stronger to B than to C.

In the following, we present an approach to flexible querying using discovered similarity knowledge represented by associative term nets. The approach was tested using an experimental prototype for querying in bibliographic databases. The databases used in the experiments were all extracted from real library databases. Most of these databases had a complex database scheme from which we extracted a subset relevant to the purpose of the investigation, as described by the following unnormalized database scheme:

```
bib_entity(Author*, Title, Year, Type, Note*, Keyword*)
```

where * indicates a multi-value field. In the following, we shall consider only bibliographic entity ('bib_entity') objects of type 'book', and only the single-value attribute Title, and the multi-value attribute Keyword. Since the keywords are chosen from a controlled vocabulary of words (terms), and attached to books by librarians, not by authors, we have a high degree of homogeneity in the keyword characterizations of objects.

Our approach to flexible querying extends conventional crisp query evaluation in two respects, namely: (1) fuzzy interpretation of single constraints (criteria): weakening such that similar values also are considered, and (2) fuzzy aggregation of single constraints in the compound query, relaxing conjunction "as much a necessary" towards disjunction. First, we introduce the associative term net derived from the database. This net is the basis for the similarity relation and thus for the fuzzy interpretation of single constraints. We then present the use of the net in query evaluation, and give examples of queries evaluated against the small database used in an experiment.

3 Discovery of term associations

The most important attribute domains in bibliographic databases are the sets of words used in different descriptions of objects in the database, e.g., in titles, abstracts, notes, related keywords, and, even, in the author names. Our main interest is the association similarity between words (or terms), and a major issue here is how to establish the term association net representing this similarity.

Since we did not have access to any useful manually created association networks, apart from the available library classification systems, and since these constitute only very sparse networks, we focused on automatically discovery of associations by, so to say, mining data in the database. For the experiment discussed in this paper, we have chosen to build a network based on librarian keywords (terms) with no influence on associations from words as used in other attributes describing objects. Thus, even titles are ignored when associating terms.

The association relation $s(\varDelta, \varDelta)$, where $\varDelta$ is the set of terms in the Keywords domain, is defined as follows,

$$s(A,B) = \begin{cases} 0 & \|\varOmega_A\| = 0 \\ \frac{\|\varOmega_A \cap \varOmega_B\|}{\|\varOmega_A\|} & \|\varOmega_A\| > 0 \end{cases} \tag{1}$$

where $\varOmega$ denotes the set of documents represented in the database, and $\varOmega_X$ denotes the subset which applies the term X in the description of the document. By this definition, the relation is reflexive, and asymmetric. Thus, in the example illustrated by Figure 2(a), we have

$$s(A,B) = \tfrac{7}{7+3} \ = 0.70$$

$$s(B,A) = \tfrac{7}{7+21} = 0.25$$

The corresponding edges in the association term net are depicted in Figure 2(b).

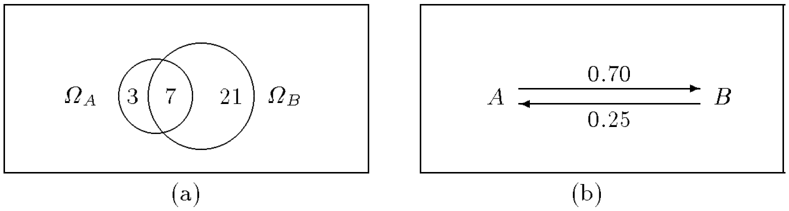

Fig. 2. Illustration of term association computation and representation: (a) frequencies of A and B, (b) resulting term associations

As an example of similarity knowledge derived in this way, consider the small section of the association term net, derived from the database in the experiment, as shown in Figure 3.

Since the approach is—as far as similarity of words used in query constraints are concerned—solely based on term associations as described above, an interesting question is: What is the kind of these associations, and how can its meaning be described? The intention is to capture relations that resemble relevant aspects of human associations. As compared to the other extreme to achieve this; knowledge acquisition by interviewing an expert in the domain in question, one major drawback appears to be the following: Words that are closely related by

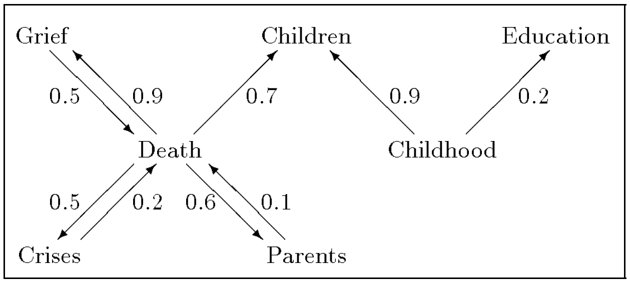

Fig. 3. A small section of the term association net derived from the database

association may also be closely related by meaning, as, for instance, 'parent' and 'mother'. When two words are closely related by meaning, a document will typically apply only one of these words, and therefore the system would not capture the association between the two words. However, since our term association net is based only on words attached from a controlled set, where each word in this set is carefully chosen to be with a distinct meaning—and thus "close meaning" is a rare phenomenon—this drawback has only little influence on the quality of the resulting association net.

Normally, we would expect an association relation to be transitive to some degree—if a house relates to a garage and a garage to a car, then we may also perceive that the house relates to the car. Even so, it appears wrong to impose transitivity on the associations obtained from mining, as suggested in our approach. Assume that there is a is a strong relation from house to garage, a strong relation from garage to car, and a weaker relation from house to car. In this situation we would expect all three relations to be reflected by the associations derived from statistics on the use of the words. Therefore, to strengthen the relationships through transitivity of the relation, is likely to lead to less intuitive results.

4 Using term associations in flexible querying

4.1 Definition of the flexible query-answer

We consider user queries expressed as a list of criteria $C_1, \ldots, C_n$ that all should be satisfied as much as possible. A criterion C_i represents a constraint on the domain of an attribute referred to by the criterion. For the experiment we considered criteria of the form *keyword* $= X$, where X is some term in the domain of keywords. Hence, queries are are expressed on the form $Q = X_1$ and $\ldots$ and X_n. Since our major objective was to explore to which extent mined knowledge may be meaningful and useful for query evaluation, we deliberately chose simple knowledge aggregation operators, such as the max operator and the arithmetic mean, without allowing for importance weighting of criteria. More sophisticated operators for query evaluation have been studied elsewhere (see, for instance, [6–8]) and leave space for further improvements.

A query is evaluated under two thresholds, δ_C and δ_Q, which delimit the extend to which a single constraint (criterion), respectively the query aggregation, can be relaxed. A constraint is relaxed through expanding the criterion term to a disjunction of all terms that the criterion term associates to, with a strength of at least δ_C. Notice, that this disjunction always contains the criterion term itself. The satisfaction of a criterion term by an object is the maximum of the satisfaction of the terms in its expansion. The overall satisfaction of the query is determined as the arithmetic mean of the satisfaction of the criterion terms; only objects that satisfy the query to at least the degree δ_Q are included in the answer. Formally, the answer to a query $Q = X_1$ and ... and X_n, with the thresholds δ_C and δ_Q, posed to a collection Ω is set defined by:

$$\text{Answer}(\Omega|Q) = \{(\omega, \sigma) \mid \omega \in \Omega,\ \sigma = \text{score}(Q|\omega),\ \sigma \geq \delta_Q\} \tag{2}$$

with

$$\text{score}(Q|\omega) = \frac{1}{n}\sum_{i=1}^{n} \gamma_i \tag{3}$$

and

$$\gamma_i = \begin{cases} 0 & \max_{Y \in \Delta_\omega} s(X_i, Y) < \delta_C \\ \max_{Y \in \Delta_\omega} s(X_i, Y) & \text{otherwise} \end{cases} \tag{4}$$

where Δ_ω is the set of terms represented as the value of the attribute Keyword of the object ω.

We notice that the max operator is just instance of the t-conorms that may be applied in Formula def:cscore. We may argue that a stronger t-conorm, such as the algebraic sum, defined by $g(a, b) = a+b-ab$ will provide a better measure. Similarly, the arithmetic mean is an instance of the averaging operators that may be applied in Formula def:qscore, and we may argue that a weaker operator, closer to the min operator, will improve the correctness of the measure. For the latter, we may apply an an OWA operator [9] with a higher degree of *andness*, say, 0.75, than the arithmetic mean which, when modeled by an OWA operator, has the *andness* 0.5.

4.2 On the effect of the thresholds on the answer

The relaxaton of the criterion terms $X_1, \ldots, X_n$ replaces each X_i by a disjunction of the set of terms which it is associated to with a strength of at least δ_C, that is, the set defined by

$$r_{s,\delta_C}(X_i) = \{(Y, \beta) \mid Y \in \Delta,\ \beta = s(X_i, Y),\ \beta \geq \delta_C\} \tag{5}$$

Thus, lowering δ_C allows more terms to be included in the set, and, thereby, more objects to satisfy the criterion. For illustration, consider the query

$$Q = \text{Death and Childhood}$$

When referring to Figure 3 we get for $\delta_C = 0.8$:

$$r_{s,0.8}(\text{Death}) = \{\,(\text{Death}, 1), \quad (\text{Grief}, 0.9)\,\}$$

$$r_{s,0.8}(\text{Childhood}) = \{\,(\text{Childhood}, 1), \quad (\text{Children}, 0.9)\,\}$$

while $\delta_C = 0.6$ leads to

$$\begin{aligned} r_{s,0.8}(\text{Death}) = \; & \{\,(\text{Death}, 1), \quad (\text{Grief}, 0.9), \\ & (\text{Children}, 0.7), \quad (\text{Parents}, 0.6)\,\} \end{aligned}$$

$$r_{s,0.8}(\text{Childhood}) = \{\,(\text{Childhood}, 1), \quad (\text{Children}, 0.9)\,\}$$

Thus, by Formula 3, an object ω with the Keyword attribute value $\Delta_\omega = \{\text{Parents, Childhood}\}$ satisfies the criterion Death to dergee 0.6 (through Parents), and the criterion Childhood to degree 1 (through Childhood). We obtain as the overall score in the query $Q =$ 'Death and Childhood' as the arithmetic mean of the two satisfactions, namely $(0.6 + 1)/2 = 0.8$. In Table 1 we show for different levels of the overall score the subsets of keywords yielding this score when applied as the value Δ_ω of the Keyword attribute of an object ω.

Table 1. Scores in $Q =$ 'Death and Childhood' by objects described by the keywords in one of the subsets

Category	Score	Values of the Keyword attribute
1	1.00	{Death, Childhood}
2	0.95	{Death, Children}, {Grief, Childhood}
3	0.90	{Grief, Children}
4	0.85	{Children, Childhood}
5	0.80	{Parents, Childhood}, {Children}
...		
N	0.50	{Death},{Childhood}
$N + 1$	0.45	{Grief}, {Children}
...		

Now, setting thresholds $\delta_Q = 1$ and $\delta_C = 1$ results in only Category 1 objects, that is, the same answer as we would obtain to the crisp Boolean query 'Death AND Childhood'. Keeping $\delta_Q = 1$ and lowering δ_C do not change anything since the score falls below 1 if just one criterion is satisfied below 1. Keeping $\delta_C = 1$ and setting $\delta_Q = 0.5$ lead to an answer consisting of Category 1 and Category N objects, that is, the same answer as we would obtain to the crisp Boolean query 'Death OR Childhood'. Setting $\delta_Q = 0.85$ and $\delta_C = 0.7$ (or lower) lead to Category 1–4 objects in the answer.

4.3 An example from flexible querying in a set of real data

To give an impression of how this works in practice, an answer to the query $Q =$ 'Death and Childhood' is shown in Table 2. The query is evaluated on a database containing about $6,000$ objects that was also the basis for the applied association term net exemplified by a small subset in Figure 3. Without relaxing the query—that is, with $\delta_Q = 1$ and $\delta_C = 1$, corresponding to the crisp Boolean query 'Death AND Childhood'—the answer is empty. For the answer shown, the chosen thresholds are $\delta_Q = 0.80$ and $\delta_C = 0.80$. Only titles and scores for the books are listed.

Table 2. Answer to 'Death and Childhood', when evaluated on the database in the experiment

Score	Title of the book
0.95	Only a broken heart knows
0.95	Baby's death—"life's always right"
0.95	We have lost a child
0.95	Loosing a child
0.95	Cot death
0.95	When parents die
0.95	To say good-bye
0.95	We have lost a child
0.95	A year without Steve
0.95	Taking leave
0.95	Grief and care in school
0.89	Can I die during the night
0.89	Children's grief
0.89	Children and grief

Where the crisp query gave an empty answer, and would have required tedious trail and error from the user's side, the flexible query evaluation gave immediately a set of obvious relevant books, in fact with a very high recall and a high precision. These results even surprised the experienced librarians associated to the project team. We should emphasize that the database used in the experiment is not a "fake". The content is descriptions of real world books on issues related to children, school and education. All descriptions are made by librarians, that have carefully examined the books and in this connection chosen distinct, describing keywords from a controlled vocabulary. The resulting association term net is in a sense really the work of the librarians. However, they are not aware that they have been doing this work. What they see is only their description. The term net however, is, so to say, a description of their descriptions.

5 Concluding remarks

We have presented ideas and preliminary results from an ongoing research project on applying fuzzy query evaluation based on domain knowledge derived from the database state. The functions used for grading objects in the answer and for building the networks are deliberately chosen as simple, partly because we want the principles to be as easily comprehensible as possible, and partly because further research has to be done to reveal properties that make functions suitable and more accurate.

In this paper we have shown only one example of a query result. However, we have developed a system, which is frequently tested by and discussed with librarians, who agree that the association based query relaxation in general leads to relevant additional objects.

The experiments performed have demonstrated that under certain assumptions on quality of data, mining for associations leads to results that support better answers to queries. Also, we consider the resulting association term net as valuable additional information, that may be of interest, not only to the casual user of the system, but also potentially to the professionals with insight knowledge of the domain, because the network constitute a kind of inversion of data and therefore may exploit patterns that are otherwise hidden even from experts.

Acknowledgment

This research was supported in part by the Danish Library Center Inc., and the Danish Natural Science Research Council (SNF).

References

1. Larsen, H.L., Andreasen, T. (eds.): *Flexible Query-Answering Systems.* Proceedings of the first workshop (FQAS'94). Datalogiske Skrifter No. 58, Roskilde University, 1995.
2. Christiansen, H., Larsen, H.L., Andreasen, T. (eds.): *Flexible Query-Answering Systems.* Proceedings of the second workshop (FQAS'94). Datalogiske Skrifter No. 62, Roskilde University, 1996.
3. Andreasen, T., Christiansen, H., Larsen T. (eds.): *Flexible Query Answering Systems.* Kluwer Aademic Publishers, Boston/Dordrecht/London, 1997.
4. Andreasen, T.: *Dynamic Conditions.* Datalogiske Skrifter, No. 50, Roskilde University, 1994.
5. Andreasen, T.: On flexible query answering from combined cooperative and fuzzy approaches. In: *Proc. 6'th IFSA, Sao Paulo, Brazil, 1995.*
6. Larsen, H.L., Yager, R.R.: The use of fuzzy relational thesauri for classificatory problem solving in information retrieval and expert systems. *IEEE J. on System, Man, and Cybernetics* **23**(1):31–41 (1993).
7. Larsen, H.L., Yager, R.R.: Query Fuzzification for Internet Information retrieval. In D. Dubois, H. Prade, R.R. Yager, Eds., *Fuzzy Information Engineering: A Guided Tour of Applications,* John Wiley & Sons, pp. 291–310, 1996.
8. Yager, R.R., Larsen, H.L.: Retrieving Information by Fuzzification of Queries. em International Journal of Intelligent Information Systems **2** (4) (1993).
9. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision making. em IEEE Transactions on Systems, Man and Cybernetics **18** (1):183–190 (1988).