Query Expansion by Semantic Modeling of Information Needs (Extended Abstract)

Piotr Wasilewski

Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw
Banacha 2, 02-097 Warsaw, Poland
piotr@mimuw.edu.pl

Keywords: information retrieval (IR), semantic information retrieval, semantic search engine, information need, semantic modeling of information need, query expansion, semantic query expansion.

1 Introduction

The paper is devoted to the semantic query expansion. It investigates the query expansion using a semantic modeling of information need. An information need is a result of information processing taking place in a user's mind. A query is a data structure expressing a given information need. Its elements can be treated as corresponding to mental representations involved in creation of the information need. On the basis of this theoretical correspondence, semantic query expansion is studied. User's information needs are semantically modeled on the basis of ontologies taken as knowledge representation systems: a query expressing a given information need is represented by a family of concepts from an ontology. Two types of ontologies are discussed: domain ontologies given by experts and automatic ontologies discovered by the IR system. Domain ontologies can be approximated using granules discovered from data, including concepts from automatically generated ontologies, but granules of other types can also be used. Various ways of query semantic modeling will be discussed. Those semantic models of queries will serve as basis for semantic query expansion. Ontologies are viewed as corresponding to conceptual hierarchies stored in users' minds. According to cognitive science, on concepts and conceptual hierarchies different operations can be performed by a human mind. On the theoretical basis of the correspondence between ontologies and user's conceptual hierarchies, different query semantic models can be discussed as corresponding to particular mental operations on concepts postulated by cognitive science.

2 Semantic Information Retrieval

Semantic retrieval is a new type of information retrieval. Information retrieval is understood as finding material (typically documents) of an unstructured nature (usually text) from large collections (usually stored on computers) that

CONCURRENCY, SPECIFICATION AND PROGRAMMING
M. Szczuka et al. (eds.): Proceedings of the international workshop CS&P 2011
September 28-30, Pułtusk, Poland, pp. 523-530

satisfies an user's information need (Manning et al., 2008). In every information retrieval system, four elements can be distinguished:

- a user's mind[1] being a source of information needs and formulated queries,
- user's interface used for entering queries
- search engine operating with an inverted index and retrieving documents,
- data repository, storing all collected documents

Information retrieval systems are typical examples of human - computer interaction systems. Traditional information retrieval system can be described as linguistic/syntactic. In such systems searching is based on the presence of words in documents. In the semantic information retrieval the meaning of words are involved, whereas searching is done by looking at the knowledge contained in documents. Thus, semantic information retrieval must be based on the some way of knowledge representation. For example, in the search engine prepared within the project SYNAT, for the purpose of knowledge representation ontologies are selected (see Wroblewska et al. 2011; Nguyen and Nguyen 2011; Ngueyn et al., 2011), they are presented as sets of concepts connected by various relations, mainly by the relation of subsumption (is-a relation), however being-a-part-of relation or other relations are also admissible (Breitman et al., 2007; Buitelaar and Cimi, eds., 2007; Colomb, 2007; Staab and Studer, eds., 2009). Additionally, we treat concepts from ontologies as meanings of words while the knowledge in ontologies is contained in relations, or also in the concepts, assuming that they are defined on the basis of attributes/slots (see Wasilewski, 2011).

In the semantic information retrieval system, a module of semantic searching is equipped with knowledge representation system, e.g. with a given ontology, while meanings are assigned to the words from document on the basis of this ontology. Therefore, in the semantic information retrieval system, meaning and further, knowledge are located in two modules of the system: in the user's mind in which they are components of information need and in the ontology incorporated in the system. In the SYNAT project it is also planned to develop user's interface to a dialogue model for user - search engine interactions, which will conduct a dialogue with the user aimed at specification of a query and driving the searching of documents or presentation of retrieved results. An important function of the module will be the translation of a query entered by the user and expressed in natural language, onto a query in an ontology based descriptive logic language. In this translation, the ontology from a semantic search engine will be also involved.

3 Semantic Modeling of Information Needs

The idea of semantic modeling of information needs was proposed in (Wasilewski, 2011). It is based on a correspondence between knowledge and meaning which are placed in users' minds on the one hand, and knowledge and meaning represented by ontologies on the other hand. In such view, ontologies, as

[1] Understood following cognitive science as an information processing system.

knowledge representation systems, correspond to conceptual hierarchies stored in human minds.

An information need arising in the user's mind consist, inter alia, of concepts. However, an information retrieval system has no access to the conceptual frames in the user's mind. Communication between the mind and the information retrieval system is done through a query formulated and entered by the user expressing his/her information need. A query is a data structure usually consisting of words. In the sequel, we assume that words contained in the query and referring to concepts (terms) are mapped by the system to concepts included in the ontology of the system. Let us note that this mapping is in fact assigning to a query its meaning in a given ontology and that the context of this ontology is essential: the same query in two different ontologies can have two different meanings. This reveals the nature of semantic modeling of queries: a given query is semantically modeled by a family of concepts interpreted as meanings of words contained in the query.

The paper will discuss various semantic models of information needs (see Wasilewski, 2011). Let $\langle O_1, \leqslant \rangle$ be an ontology used by a given semantic search engine[2]. Hereafter we model the information need semantically as a set of concepts from ontology O_1: $\{C_1, ..., C_n\} \subseteq O_1$ determined in some way by query q expressing this information need. Such family we will call *a semantic model of query q* or *a semantic model of information need*. Because information need is always expressed in the form of a query, we will also briefly say that the family of concepts models semantically the query.

Six query sematic models will be discussed: three simple and three complex. Semantic modeling depends on ontologies as well as on algorithmic methods of assigning concepts to documents (conceptual indexing algorithms). The first query semantic model given below depends essentially only on conceptual indexing, a given ontology only narrow a scope of concepts which can be assigned to query. In this model a structure of the ontology O_1, representing knowledge, is not involved. From that point of view, among query semantic models given below, the first query one can be view as basic, while the next five as derivable from it on the basis of knowledge represented by ontology O_1.

Simple Query Semantic Models

1. The simplest way of semantic modeling of the information need expressed by query q is to take concepts from the ontology: O_1 which are assigned by the system to terms contained in query q. Such family of concepts we will denote by $O_1(q)$.
2. Another way of modeling query q is to take additionally concepts from ontology O_1 which are placed between concepts from family $O_1(q)$ which are comparable with respect to subsumption relation $\leqslant$. Such family we will

[2] Here, we adopt simplifying assumption about the ontology: an ontology is understood as a set of concepts O_1 partially ordered by subsumption relation $\leqslant$: $\langle O_1, \leqslant \rangle$. If it will not lead to confusion (a subsumption relation will be understood from the context), to ontology $\langle O_1, \leqslant \rangle$, as a partial order, we will refer also by O_1.

denote by $O_1[q]$, in other words:

$$O_1[q] := \{D | \exists A, B \in O_1(q); A \leqslant D \leqslant B\}. \tag{1}$$

Let us note that family $O_1[q]$ can be empty even when $O_1(q)$ is nonempty, and this is when $O_1(q)$ is an anti-chain, i.e. any two concepts from $O_1(q)$ are not comparable with respect to subsumption relation $\leqslant$. Taking into account such possibility, we can introduce next ways of sematic modeling of information needs.

3. Query q can be also modeled by the family of all concepts from ontology O_1 which ar comparable by the subsumption relation with at least one concept from family O_1, i.e. take the family of the form:

$$FI_{O_1(q)} = \bigcup_{A \in O_1} (A] \cup \bigcup_{A \in O_1} [A), \tag{2}$$

where $(A]$ and $[A)$ are respectively a principal filter and a principal ideal determined by concept A in partially ordered set $\langle O_1, \leqslant \rangle$. As versions of this model, families $(A]$ and $[A)$ can be taken separately.

Complex Query Semantic Models[3]

4. Family $O_1(q)$ can be taken as a set of generators of a complete lattice: we take family $O_1(q) \subseteq O_1$ as partially ordered set $\langle O_1(q), \leqslant_{|\langle O_1(q) \rangle} \rangle$ and then we take the Dedekind-MacNeille completion of $\langle O_1(q), \leqslant_{|\langle O_1(q) \rangle} \rangle$ which is a complete lattice[4] For the family semantically modeling query q we take the universe of this lattice denoted by $L[O_1(q)]$.
5. Let us note that family $L[O_1(q)]$ not necessarily contains e.g. all upper bounds of family $O_1(q)$ in set $\langle O_1, \leqslant \rangle$ (upper bounds of family $O_1(q)$ are superconcepts of all concepts from family $O_1(q)$)[5]. In order to consider all elements somehow generated from family $O_1(q)$ we can proceed in two ways. Firstly, the Dedekind-MacNeille completion of whole ontology O_1 is taken, denote the universe of this lattice by $L[O_1]$ (note that $O_1 \subseteq L[O_1]$. Then take family $O_1(q)$ as a set of complete generators and generate complete sublattice $Sg_{L[O_1]}(O_1(q))$ of the complete lattice $L[O_1]$. For the family semantically modeling query q we take family $Sg_{L[O_1]}(O_1(q))$.
6. Secondly, take the family $FI_{O_1(q)}$ and then take the Dedekind-MacNeille completion of partially ordered set $\langle FI_{O_1(q)}, \leqslant_{|FI_{O_1(q)}} \rangle$, the universe of this complete lattice will be denoted by $L[FI_{O_1(q)}]$. For the family semantically modeling query q we take family $L[FI_{O_1(q)}]$.

[3] These query semantic models are complex in the sense that they are various completions of first three simple query semantic models by means of partially ordered set operations as well as algebraic operations.

[4] On of the methods of construction of the Dedekind-MacNeille completion is creating a concept lattice (Wille, 1982; Ganter and Wille, 1999) for a given partially ordered set (see Dedekind completion theorem in Ganter and Wille, 1999). Creating finite concept lattices has a computational character.

[5] All lower bounds of family

Note that two last methods of modeling of information needs outlined above are different and have their own advantages and disadvantages. Lattices $Sg_{L[O_1]}(O_1(q))$ and $L[FI_{O_1}]$ do not have to be isomorphic.

4 Query Expansion

Among key notions in information retrieval are *information need* and *relevance of a document*. From the very beginning of the notion of information need, it was highlighted that information need has both conscious and unconscious components: it is a desire of an individual person or group of people to find and get information satisfying their conscious or unconscious needs (demands) (Taylor, 1967). In other words, an information need is a topic on which a user would like to know more, and it is distinguished from the query - a data structure which is entered to a IR system by a user in order to communicate information need (Manning et al., 2008). *Relevance* indicates how well a document or set of documents satisfies the user's information needs (Cuadra and Katter, 1967). In other words, the document is *relevant* if it is perceived by the user as containing valuable information with regard to its information needs (Manning et al., 2008). Relevance is traditionally of binary nature: the document is relevant or irrelevant (Butcher et al., 2010; Manning et al., 2008), however at the beginning of information retrieval evaluation, the first Cranfield experiments used a five-point scale of relevance (Cleverdon, 1967; Vorhees, 2002; Voorhees and Harman, eds., 2005). Recently, *graded relevance* again become used in the evaluation experiments (Najork et al., 2007; Butcher et al., 2010).

A query is a data structure expressing a given information need. An information need appears in a user's mind as a result of information processing which consists of both conscious and unconscious components. This process can lead to a query formulation[6]. Query expansion is a process of query reformulating aimed at improving the search results by expanding the search query to retrieve additional documents. Such expanding is made using new topics somehow connected to those contained in the original search query. This rests on the assumption that queries are not formulated with the best words or that they are expressed using too few words.

In the process of information need formulation, knowledge stored in the user's mind as well as results of perception, including understanding of communication, are involved. According to the computational-representational understanding of mind, the most common approach in the contemporary cognitive science (Thagard, 2005), stages of this process consist of mental representations, also knowledge and perception results are build of those mental representations. The formulation of an information need can be viewed as an interaction process between, among others, knowledge stored in a user's mind and results of the user's communication with the environment, including information coming from text reading. The expression of an information need by a query can be supported by

[6] It is not necessary that a given information need is expressed by a query and can be expressed by various queries

means of a dialog of a search engine with a user. The information need creation process can be treated as an internal interaction (Skowron, Wasilewski, 2011), i.e. interaction of a user's mind, understood as an information processing system, with the internal environment. A dialog with an IR system is an example of interaction of a user's mind with the external environment. A similar dialog can also support a query expansion, in this case it is an example of interaction of an IR systems with the external environment, containing a user. Therefore, a query formulation as well as a query expansion supported by a dialog can be treated as highly-interactive processes (Skowron, Wasilewski, 2011).

The first semantic query expansion was proposed by Ellen M. Vorhees (1994). Queries were expanded by means of synonymy and hyponymy/hypernymy (is-a-relation) relations within WordNet. It was shown that such query expansion made a little difference in retrieval results when original queries were relatively complete descriptions of information that was sought while less developed queries (usually consisting of a single sentence) were significantly improved (Vorhees, 1994).

The paper will discuss semantic query expansion made by means of semantic modeling of information needs. It is based on the correspondence between mental representations in user's mind of which information needs are built and knowledge and concepts contained in ontologies. In this approach an original search query is expanded by concepts from semantic models derivable from its basic semantic model.

Cognitive science postulates that on concepts and conceptual hierarchies in human minds various operations can be performed, including *inheritance*, *generalization*, *spreading activation* and *inferences*. Since ontologies are treated as corresponding to conceptual hierarchies in users' minds, then various query semantic models can be discussed as corresponding to particular mental operations. For example, families $\bigcup_{A \in O_1}(A]$ and $\bigcup_{A \in O_1}[A)$, taken as query semantic models, reflect mental operations inheritance and generalization respectively while complex query semantic models can be treated as corresponding to inferences.

Derivable query semantic models, especially complex models, can be too broad for query expansion. In such case instead of algebraic generations in the case of complex models, induction procedure can be considered:

$$E_{i+1}(\varPhi) := \{f(A, B) : A, B \in E_i(\varPhi)\ \mathit{and}\ f \in \{\wedge, \vee\}\}, \tag{3}$$

where $E_0(\varPhi) = \varPhi$ and for family $\varPhi$ families $L[O_1(q)]$, $Sg_{L[O_1]}(O_1(q))$ or $L[FI_{O_1(q)}]$ can be taken. Similarly, from families $\bigcup_{A \in O_1}(A]$ and $\bigcup_{A \in O_1}[A)$ chains starting at A and having the length 1,2,3... can be selected. The paper will discuss such possibilities and particular examples will be presented.

Acknowledgements: Author would like to thank Andrzej Skowron, Hung Son Nguyen, Dominik Ślęzak, Wojciech Jaworski, Wojciech Świeboda and other colleagues from the SYNAT group for their critical comments and valuable discussions helping to improve the paper. Research was supported by the National Centre for Research and Development (NCBiR) under the grant SP/I/1/77065/10 by the Strategic scientific research and experimental development program: ”Interdisciplinary System for Interactive Scientific and Scientific-Technical Information” and by the Foundation for Polish Science under the individual research grant by the program Homing Plus, edition 3/2011.

References

1. Akrivas, G., Wallace, M., Andreou, G., Stamou, G., Kollias, S.: Context-Sensitive Semantic Query Expansion. In: Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS'02). IEEE Computer Society, Washington (2002) 109–114.
2. Berry, M.W., Browne, M.: Understanding Search Engines. Society for Industrial and Applied Mathematics (2005).
3. Breitman, K.K., Casanoca, M.A., Truszkowski, W.: Semantic Web: Concepts, Technologies and Applications. Springer Verlag (2007).
4. Buitelaar, P., Ciminao, Ph. (Eds.): Ontology Learning and Population: Bridging the gap between Text and Knowledge. IOS Press (2008).
5. Butcher, S., Clarke, Ch.L.A., Cormack, G.V.: Information Retrieval: Implementing and Evaluating Search Engines, Massachusetts Institute of Technology Press (2010).
6. Cleverdon, C.W.: The Cranfield tests on index language devices. In: Aslib Proceedings, volume 19, (1967) 173-192. (Reprinted in Readings in Information Retrieval, K. Sparck-Jones and P. Willett, editors, Morgan Kaufmann, 1997).
7. Colomb, R.M.: Ontology and the Semantic Web. IOS Press (2007).
8. Cuadra, C.A., Katter, R.V.: Opening the black box of relevance. Journal of Documentation **231:4** (1967) 291–303.
9. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer-Verlag (1999).
10. Jankowski, J., Skowron, A.: Wisdom Technology: A Rough-Granular Approach. In: M. Marciniak, A. Mykowiecka (Eds.), Bolc Festschrift. Lectures Notes in Computer Science. **5070** (2009) 3–41.
11. Jankowski, J., Skowron, A.: Rough Granural Computing in Human-Centric Information Processing. In: K.A. Cyran, S. Kozielski, J.F. Peters, U. Stańczyk, A. Wakulicz-Deja (Eds.), Man-machine interactions. Springer (2009) 23-42.
12. Jarvelin, K., Kekalainen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20:4** (2002) 422–446.
13. Manning, Chr.D., Raghavan, P., Schutze, H.: Intorduction to Information Retrieval, Cambridge University Press (2008).
14. Najork, M.A., Zaragoza, H., Taylor, M.J.: HITS on the Web: How does it compare. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2007) 471–478.
15. Nguyen, H.S., Ślęzak, D., Skowron. A., Bazan, J.G.: Semantic Search and Analytics over Large Repository of Scientific Articles. In: Proceedings of the SYNAT Workshops. Studies in Computational Intelligence. Springer Verlag (2011).
16. Nguyen, L.A., Nguyen, H.S.: On Designing the System SYNAT. In: Proceedings of the SYNAT Workshops. Studies in Computational Intelligence. Springer Verlag (2011).

17. Robertson, S.: On GMAP and other transformations. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management (2006) 78–83.
18. Skowron, A., Szczuka, M.: Toward interactive computations: A rough-granular approach. In: J. Koronacki, S. Wierzchon, Z. Ras, J. Kacprzyk (Eds.), Commemorative Volume to Honor Ryszard Michalski. Springer-Verlag (2009) 1-20.
19. Skowron, A., Wasilewski, P.: An introduction to perception based computing. In: T.-h. Kim et al. (Eds.): FGIT 2010, Lecture Notes in Computer Science **6485** (2010) pp. 1225.
20. Skowron, A., Wasilewski, P.: Information Systems in Modeling Interactive Computations on Granules, Theoretical Computer Science, doi: 10.1016/j.tcs.2011.05.045 (2011).
21. Staab, S., Studer, R. (Eds.): Handbook on Ontologies. Springer Verlag (2009).
22. Taylor, R.S.: Process of Asking Questions. American Documentation **13** (1967) 291–303.
23. Thagard, P.: Mind: Introduction to Cognitive Science. (2nd ed.) MIT Press (2005).
24. van Rijsbergen, C.J.: Information Retrieval, 2 edition, chapter 7, Butterworths (1979).
25. Vorhees, E.M.: Query expansion using lexical-semantic relations. In: W. Bruce Croft and C. J. van Rijsbergen (Eds.). Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94). Springer (1994) 61–69.
26. Vorhees, E.M.: The philosophy of information retrieval evaluation. In: C.A. Peters et al. (Eds.): Proceedings of the Cross Language Evaluation Forum (CLEF) of 2001, Lectures Notes in Computer Science **2406** (2002) 355–370.
27. Voorhees, E.M., Harman, D.K. (Eds.): TREC. Experiment and Evaluation in Information Retrieval. Massachusetts Institute of Technology Press (2005).
28. Wasilewski, P.: Towards semantic evaluation of information retrieval. In: Proceedings of the SYNAT Workshops. Studies in Computational Intelligence. Springer Verlag (2011).
29. Wille, R.: Restructuring lattice theory. In: I. Rival (Ed.) Ordered Sets. Reidel (1982).
30. Wróblewska, A., Podsiadły-Marczykowska, T., Bembenik, R., Protaziuk, G., Rybiński, H.: Methods and tools for ontology building, learning and integration application in the SYNAT project. In: Proceedings of the SYNAT Workshops. Studies in Computational Intelligence. Springer Verlag (2011).