QUERY EXPANSION IN PERSONAL QUERIES

Indra Seher
School of Computing and Mathematics, College of Health & Science
University of Western Sydney, NSW, Australia.

ABSTRACT

There are many query expansion techniques that are being developed and used to expand user queries to retrieve more relevant Web documents, improving search results. Some of these approaches add related terms to the query by accepting the terms from the user (interactive query expansion-IQE), and in some other approaches, retrieval systems add expansion terms using various domain independent and dependent sources(automatic query expansion-AQE). This paper is based on a research which proposes a domain independent approach for AQE for personal queries, using contextual information. Compared to general queries such as definitional and descriptional queries, expansions to personal queries can be much improved with user preferences in the domain of the query. If an automated system could keep or capture the contextual information related to the query, then it could expand the query with domain specific information and with user preferences related to the query, before processing it. This paper describes the stages necessary for such an expansion and the contextual information required for each of the stages.

KEYWORDS

Question answering systems, Context model.

1. INTRODUCTION

There are many Question Answering(QA) systems that are being developed and used. Some of them are closed domain(Gandhe, Gordon et al. 2006; Goto, Miyazaki et al. 2006) and some are open domain (Cardie, Ng et al. 2000; Moldovan, Pa et al. 2003; Mori, Nozawa et al. 2005). Some of them, especially the open domain question answering systems have distinct question taxonomies and answer specific types of questions such as definitional, spatial, and temporal. Most question answering systems use a variety of linguistic resources to help in understanding the user's query and matching sections in Web documents((Pasca 2004; Isozaki 2005). Some of the QA systems use query expansion techniques to answer the questions.

Query expansion techniques aim to improve user's search by adding new query terms to an existing query(Ruthven 2003; Kelly, Dollu et al. 2005). In some, expansion terms are added to the query by the user (interactive query expansion-IQE), and in some others, expansion terms are added by the retrieval system (automatic query expansion-AQE). AQE techniques have a significant advantage over IQE such as relevance feedback, because they require no effort on the part of the user. Most of the AQE techniques apply various methods to the Web documents retrieved using terms in the query(Carpineto, Mori et al. 2001; Mano and Ogawa 2001). Some of the AQE techniques use contextual information such as the initial user profile that the user inputs, the event viewing history behaviour of the user, and past user tasks, to expand user query(Appan and Sundaram 2004; Wen, Lao et al. 2004). Some of these AQE systems are domain specific. This research considers a domain independent approach for AQE for personal queries, using contextual information such as location of the user, and short term and long term preferences of the user.

Compared to general queries such as definitional and descriptional queries, results of the personal queries can be much improved, if the receiver knows user context, and could make use of them to process the query. When humans communicate, a personal query made by one human, often do not completely describe the requirements of the human. But, still the other human involved in the communication can understand the messages better due to the richness of the languages and when they share the same context. Even if a personal query describes the requirements of the user, if the receiver knows user preferences related to the query, more useful answers can be given back to the user.

Similar to the human communication, personal queries can be understood better by an automated system, if the system can identify the query domain, and knows the user preferences that are related to the domain of the query. Personal queries could contain short term user preferences specified directly or indirectly and could also be incomplete. If an automated system can identify these long term, short term, and missing preferences of the user that are necessary and useful to process the query, and if it could keep or capture these information, then it could expand the query with such information before processing it. This paper is based on a research related to these factors, understanding user requirements from the query by identifying the query domain and user preferences in the domain, and expanding user query using contextual information.

2. TRAINING QUERIES

To understand and capture the nature of user queries, about 250 queries were collected from eight different users with different backgrounds such as administrative assistant to PhD research student in IT. The users were given with six selected application domains, and were requested to give their queries similar to the way they ask it from a friend. The domains used were requesting a meal, looking for a house to purchase/lease, getting a course information to follow a course, getting an appointment time with a Doctor, getting flight details for the purpose of traveling, and requesting to settle a bill. These queries were found to be useful in identifying the nature of user queries in specific application domains.

3. STAGES OF THE QUERY EXPANSION PROCESS

The stages given in figure 1 are identified as necessary for the expansion of a personal query. The contextual information identified to be necessary for the expansion is stored in the context model which includes the domain model, user model, spatial model and temporal model. The domain model will have a similar, well defined structure for all domains. But, before using the system in a new domain, the domain information has to be added to the existing domain model. The spatial and the temporal models are domain independent. The user model will have long term preferences of the user. Some of these preferences are domain independent, such as personal details of the user. For each of the domains implemented, user's own preferences that are specific to the domain have to be added to the user model. In order to do this, when a domain is added to the system, a domain template consisting possible types of preferences in the domain has to be added as part of the domain model.

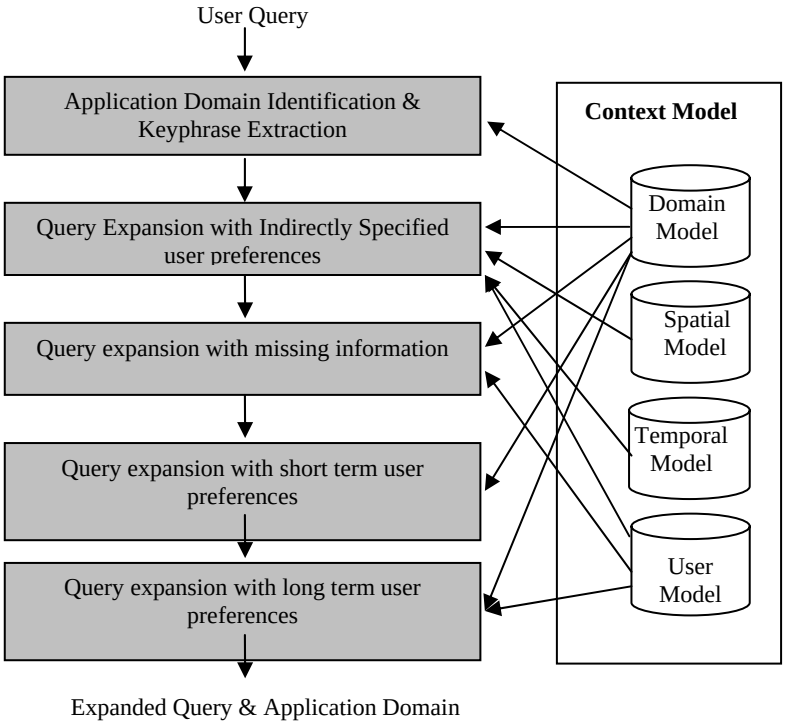

Figure 1. Stages of the query expansion process

3.1 Application domain identification and keyphrase extraction.

When a query is given by the user, the application domain of the user query has to be initially identified. This application domain is necessary for the following stages which are extracting meaningful terms from the query, identifying the intended meanings of important terms in the query, identifying whether the user has directly or indirectly specified preferences, and identifying whether all information necessary to process the query are given or missing in the query.

User queries in any application domain request information about concepts such as person, place, thing, service or provider in the domain. Such concepts are entities in the domain and have properties describing them. A domain model containing the entities, their relationships, attributes and instances of the domain will be used to identify useful terms in the query and the context of the terms. The domain model is a static model, which has to be created manually. Before using the system to a domain, the domain has to be analysed manually, to identify the concepts that can be used in the model. As an entity, a property and an instance can belong to more than one domain, entities and attributes are categorized based on their significance in the domain, and weighting methods will be used by the system to choose the most suitable domain, out of the domains identified.

Significance of an entity is decided based on whether the presence of the entity can be used to identify the domain, in which case the entity is referred to as an Identifying Entity, in this paper. For example, in the query *List me the meals in Chinese Dragon*, the term *meal* in the query is an entity of the application domain requesting meal and can be used to identify the application domain. If the presence of an entity cannot be used to identify the domain, but still the entity describes the domain, then it is referred to as a Non-Identifying entity. For example, in the query *List the flights available to any Asian country*, the term *country* in the query is an entity of the application domain flight information, but it cannot be used to identify the application domain, but still describes the domain. Similarly, significance of a property is decided based on whether the presence of the property can be used to identify the entity. For example, in the query *List me the pizzas from Dominos*, the term *pizza* is a property *meal name* of the entity *meal*, and could be used identify the entity(meal). The hierarchy of the entities, properties and instances in the domain model is shown in figure 2. Here, instances are long lists, and can be maintained separately. This will allow the instances to be reused by more than one domain.

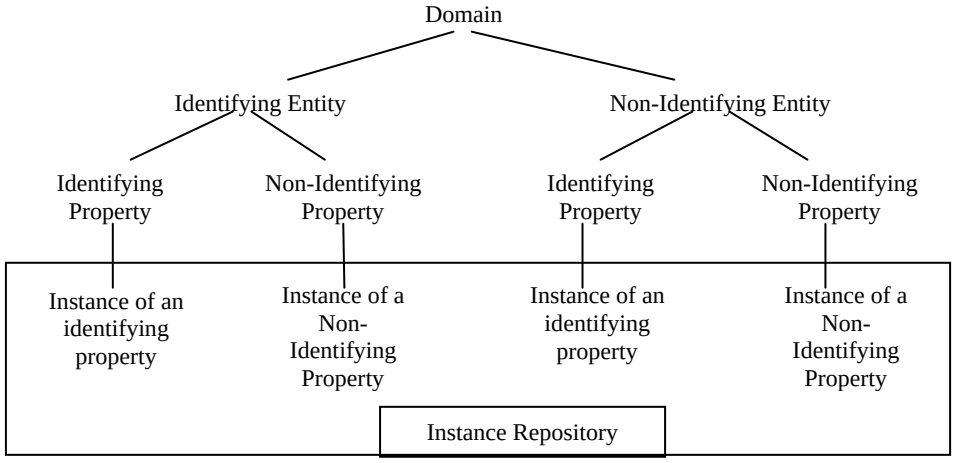

Figure 2. Hierarchy of entities, properties and instances in a domain

The weighting methods decide the context of extracted terms, based on the other terms extracted. For example, in the query *List me the pizzas from Dominos*, terms in the query that could be used to identify the application domain are *Pizza* - which is name of a meal, and *Dominos* - which can be identified as a restaurant or a game. But, when the relationships among the types of extracted terms are considered, *meal* has a relationship with *restaurant* in the *meal* domain, and hence *Dominos* has to be a restaurant name compared to a game name.

3.2 Query expansion with indirectly specified user preferences.

Often, queries do contain user preferences. Some of these preferences can be indirectly specified in the query. This is because users assume that the receivers know their preferences. For example, in the query *List me the closest McDonalds*, user preference is the location of the restaurant, which is specified indirectly by the term *closest*. Similarly in the query, *Get me appointment times with my Doctor*, *my Doctor* is an indirect preference of the user. These preferences can be personal, spatial, or temporal. Personal preferences can be general or domain specific. To add these preferences to the expanded query, user model containing user preferences has to be used. Spatial and temporal terms used to specify preferences indirectly, are always relative(*List me the closest McDonalds)*. If they are given with absolute values in the query (*List me the McDonalds in Parramatta* where *Parramatta* is a suburb), then the preference is directly specified.

3.3 Query expansion with missing information.

Some application domains require compulsory information to process queries. For example, to process a query in the flight booking domain, the start and the destination locations should be given by the user. But, often there can be questions such as *List the flight times to China in early December*, where the starting location is not given by the user. These kinds of information that are compulsory to process the query are referred to as missing information in this paper. In order to identify missing information in the query, the domain model should have the compulsory information required for processing, if there are any, for each implemented domain. When the system identifies that the compulsory information necessary for processing a query in the query domain is not specified in the query, it could add them from user model, which will have general user information/preferences and preferences of the user in the specific domain.

3.4 Query expansion with short term user preferences.

Though user's could have their preferences stored as part of their profile in the user model, the preferences can vary at times. The most recent preferences are issued as part of user query, which have more value than stored user preferences. Some preferences that are not stored as part of user preferences, can also be issued as part of user query. In the query *List me the McDonalds in Parramatta*, a user preference which is the location of the restaurant, is given in the query. Such types of preferences of domains have to be maintained as part of the domain model to identify them in the user query.

3.5 Query expansion with long term user preferences.

Even if a query contains all information necessary to process the query, if user preferences related to the query are available, these preferences could also be used to process the query. When these preferences which are supplementary information, are available, more and useful expansion can be done to the user query. This will help to narrow down the searches, which in turn will use more relevant data sources, resulting more accurate results. For example consider the query *List meals available at Chinese Dragon*. Though this query is complete, and can be used for processing, if user's preferred meals, preferred cost, preferred location etc. are known, a better result can be produced.

4. BUILDING THE CONTEXT MODEL

Based on the stages of the query expansion process, and the contextual information required for each of the stages, to use the system for a domain, initially the domain has to be analysed manually to identify the concepts that could be used in the domain model. This will include entities in the domain, categories of the entities as identifying/non-identifying, relationships among the entities in the domain, properties, their categories as identifying/non-identifying and instances of the properties in the domain. These should be added to the domain model. If the domain requires any compulsory information to process queries, they also have to be included in the domain model. Since the types of preferences in a domain are required for the

expansion, a template consisting possible preferences in the domain has to be identified manually, and stored as part of the domain model. The spatial model will have spatial references which could be absolute or relative. The absolute spatial values can be organized in a hierarchy. The temporal model also could have absolute and relative values. For a new user to use the system, user model consisting user information and long term preferences of the user has to be updated by adding user information. Some of these preferences are domain independent(personal details of the user), and the rest are domain specific(user preferred meal in the domain meal, user preferred flight classes in getting flight information). The domain template with types of preferences in the domain which is stored as part of the domain model, has to be filled with the preferences of the user and should be stored as part of user model.

5. CONCLUSION

This paper describes an ongoing project on how different contextual information can be used to expand user queries to answer personal queries. The main steps necessary for the expansion and the required components of the context model are explained. Currently the domain model and a weighting algorithm are designed to identify the application domain, which is the key step in the expansion process. As many of the instances of the properties are proper nouns, some of the most popular thesauruses are being considered to see the possibilities of making use of them in this system. The system will be evaluated to utilise and finalise the algorithm.

REFERENCES

Appan, P. and H. Sundaram, 2004. Networked multimedia event exploration. *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*. ACM Press, pp 40--47

Cardie, C., V. Ng, et al.,2000. Examining the role of statistical and linguistic knowledge sources in a general-knowledge question-answering system. *Proceedings of the sixth conference on Applied natural language processing*, Seattle, Washington, pp. 180--187.

Carpineto, C., R. d. Mori, et al., 2001. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, Vol. 19, No. 1, pp 1--27.

Gandhe, S., A. S. Gordon, et al.,2006. Improving question-answering with linking dialogues. *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*, Sydney, Australia, pp. 369--371.

Goto, J., M. Miyazaki, et al.,2006. A TV agent system that integrates knowledge and answers users' questions. *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*, Sydney, Australia, pp. 300--302.

Isozaki, H., 2005. An analysis of a high-performance Japanese question answering system. *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 4, No. 3, pp 263--279.

Kelly, D., V. D. Dollu, et al.,2005. The Loquacious User: A Document-Independent Source of Terms for Query Expansion. *SIGIR'05*, Brazil, pp. 8.

Mano, H. and Y. Ogawa,2001. Selecting expansion terms in automatic query expansion. *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States, pp. 390--391.

Moldovan, D., M. Pa, et al., 2003. Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inf. Syst.*, Vol. 21, No. 2, pp 133--154.

Mori, T., M. Nozawa, et al., 2005. Multi-answer-focused multi-document summarization using a question-answering engine. *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 4, No. 3, pp 305--320.

Pasca, M., 2004. Acquisition of categorized named entities for web search. *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM Press, pp 137--145

Ruthven, I., 2003. Re-examining the potential effectiveness of interactive query expansion. *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM Press, pp 213--220

Wen, J.-R., N. Lao, et al., 2004. Probabilistic model for contextual retrieval. *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, pp 57--63