IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB12734

This paper appears in the book, *Emerging Trends and Challenges in Information Technology Management, Volume 1 and Volume 2* edited by Mehdi Khosrow-Pour © 2006, Idea Group Inc.

A Framework for Context-Aware Question Answering System of the Math Forum Digital Library

Shanshan Ma & Il-Yeol Song
College of Information Science & Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, {shanshan.ma, song}@drexel.edu

ABSTRACT

Most existing question answering systems do not always provide satisfying answers to questioners. One way to improve this situation is to incorporate contextual information into answering systems. We present a framework for a multi-dimensional (MD) **Con**text-aware **Q**uestion **A**nswering **S**ystem (**ConQAS**) for the Math Forum digital library. Our MD context model uses the following four context dimensions: the user's navigation path, the domain specific ontology term representing the question, the question type, and the user level. The model would provide more accurate answers by using the contextual information. The MD context-aware component also reuses the questions and answers from the archives.

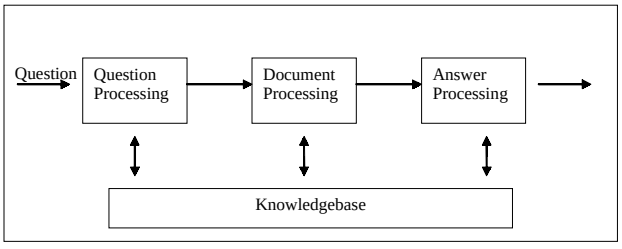
INTRODUCTION

The task of a question/answering (Q/A) system is to find answer to a question by searching a collection of documents and return only the most relevant answers. The typical work flow of a Q/A system is shown in Figure 1.

Currently most Q/A systems extract answers based on information from questions without considering any contextual information of the user. However, contextual information is very important in answering questions. Consider the following question: Where is the Empire State Building? The appropriate answer depends on the context in which the question is asked. If the question is asked on 33rd street in Manhattan, the answer might be "just around the corner on your right." If it is asked in San Francisco, the answer might be "In Manhattan, New York City." Or if it is asked in Beijing, China, the answer might be "In New York City, USA."

Answers could be more specific and accurate when we use contextual information, which is the location of the questioner in the above example. Although contextual information is easy to sense and use in a physical environment with human intermediaries, it is hard to attain when questions are answered by question answering systems. Also, what kind of contextual information should be used in system? How much contextual information should be incorporated in the system? How to find a way to make the best of contextual information for the Q/A system, in our particular situation? All these questions motivated us to propose a context-aware question answering system. Our paper will present a framework for a **Con**text-aware **Q**uestion **A**nswering **S**ystem (**ConQAS**) for MathForum Digital Library.

Figure 1. Data flow of a regular Q/A system



The MathForum (www.mathforum.org) is one of the most widely used online mathematics education services, with over one million pages of content. We work with the *Ask Dr. Math* service of the MathForum. *Ask Dr. Math* receives about 500,000 visits per month and visitors ask about 9000 questions per month. Over 200 trained volunteers, called *math doctors*, answer as many of these questions as possible. About 8000 questions and answers are in its searchable archives. Approximately half of the questions receive answers and the others go unanswered. Approximately 70% of the users who receive answers to their questions search the archive before posing the question. The large number of visits, questions, unanswered questions, and the large searchable archive make this an ideal environment for developing **ConQAS**.

The rest of the paper is organized as follows: Section 2 introduces some related work, and Section 3 describes the architecture of **ConQAS**. Section 4 presents our context model, and Section 5 presents the process of **ConQAS**. Section 6 concludes the paper.

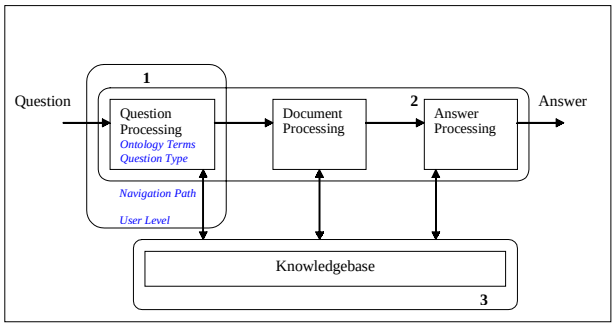
RELATED WORK

Contextual information is difficult to obtain and process. Korkea-aho (2000) summarized four main context supportive frameworks: Stick-e Notes Framework, Context Toolkit, Situated Computing Service, and Virtual Information Towers. Dey and Abowd developed a context-aware system using Toolkit (Dey& Abowd, 2000) for supporting reminders, called Cybreminder. Their context-aware system uses context for specifying reminders, beyond simple time and location and for proactively determining when to deliver them. It allows users and third parties to submit reminders. It creates reminders using a variety of input devices and receives reminders using a variety of devices, appropriate to the user's situation.

Adomavicius et al. (2005) proposed a multidimensional context model for a movie recommendation system. By using the multi-dimensional approach, the recommendation system can provide recommendations based on additional contextual information, besides the typical information on users and items. The additional contextual information here includes the place that the movie can be seen, time when the movie can be seen, and companion with whom the movie can be seen.

In the field of question answering system, QUALIFIER is a so-called event-based question answering system, featuring that it uses external knowledge resources to meet the gap between queries and documents (Yang et al., 2003). HITIQA is a scenario-based question answering system, featuring that the system would obtain new information about the user during the interaction process between the system and the user

Figure 2. Data flow of ConQAS



(Small et al., 2004). Vicedo et al. (2001) discuss a semantic-based question answering system. The semantic-based approach accomplishes question representation by combing keywords with a semantic representation of expected answer characteristics. Answer ranking is performed by computing a similarity between this representation and documents sentences. However, none of these systems fully consider important roles of contextual information in a question answering system.

SYSTEM ARCHITECTURE

ConQAS contains three modules: Context Manager, Knowledge Manager, and Retrieval Manager. The function of each module is as follows.

- **Context Manager** (The frame labeled by 1): It is the context aware module for ConQAS. Context Manager collects contextual information from four dimensions: the navigation path, the domain specific ontology term, the question type and the user level. Context Manager uses contextual information in a multi-dimensional way.
- **Knowledge Manager** (The frame labeled by 2): It maintains a knowledgebase, which is hierarchically structured by domain experts. Since most reference services encompass multiple domains, the knowledge base may involve unrelated subjects. The knowledge base will also store questions and their related answers with their context information.
- **Retrieval Manager** (The frame labeled by 3): It takes a query, analyzes the question, searches the knowledge base, ranks answers, and displays the answers in the ranked order. The question analysis function of Retrieval Manager provides Context manager with three dimensions of contextual information: the ontology terms, the question type and the user level.

THE MULTI-DIMENSIONAL MODEL FOR CONQAS

General Description of the MD Model

Accurate detection of answers from a knowledgebase depends on the degree to which relevant contextual information is incorporated in our context aware module. It is important to incorporate the contextual information of the user's scenario when he asks questions to **ConQAS**. To provide accurate answers incorporating contextual information from different perspectives, we use a multi-dimensional model (Adomavicius, et al., 2005). We consider four dimensions in our multi-dimensional model, including the user's navigation path, the domain specific ontology term, and the question type, and the user level.

Formally, let D1, D2, …, Dn be dimensions, each dimension Di being a subset of a Cartesian product of some attributes (or fields) Aij, (j = 1, …, k), i.e., Di ⊆Ai1* Ai2 * … *Aik, where each attribute defines a set of values. One or several attributes form a key: they uniquely define the rest of the attributes. For example, in our case, there are two attributes for the dimension of navigation path, which are visited nodes and time spent on these nodes.

- Navigation path dimension (D1) is defined as Navigation Path ⊆ Nodes * Time and consists of a set of paths defined by a sequence of nodes and time spent on them.
- The dimension of domain specific ontology term (D2) is defined as Domain Specific Ontology Term⊆Term 1* Term 2* …Term n, and consists of a set of ontology terms representing the question.
- Question type dimension (D3) is defined as Question type ⊆Question types and consists of different question type names. The question type is used to determine the expected answer type.
- User level dimension (D4) is defined as User Level⊆User Levels and consists of different levels;

Given these four dimensions, we define a context space for the question answering system: S = D1 * D2 * D3* D4. The answer to the question could be defined over the space D1 * D2 * D3 * D4 as Answer: D1 * D2 * D3 * D4 Answer. The answer on the context space S = D1 * D2 * D3 * D4 can be viewed in a multidimensional cube, such as the one shown below.

The First Dimension: Navigation Path

The navigation path (Shahabi et al., 1997) is one dimension of the profiler of the user who uses **ConQAS**. We know that approximately 70% of users who receive answers to their questions search the archive before posing the question in our working environment. As a user navigates through the knowledge base of our question answering system, the history of visited nodes is captured.

A navigation path could be shown as a sequence of visited nodes. The nodes here are the topics and subtopics of the hierarchically structured knowledgebase. Another important feature of the user' navigation path is the time they spent on different pages or nodes. This is the time that the user spent viewing the page, excluding the time spent on receiving and loading the page. We show the node sequence of the path traversed by the user u as Su = (N1, N2… Nn), the corresponding visiting time sequence as Tu = (Tu(N1),…, Tu(Nn)). Nodes and time are the two attributes for this dimension: D1 = Nodes* Time.

Therefore, the navigation path can be determined by the node sequence and the corresponding time. For example, an example of a navigation path on the MathForum site is as follows:

Homepage: 10 seconds → Ask Dr. Math: 8 seconds → Middle schools: 25 seconds → Number sense/ about numbers: 5 seconds → Middle schools: 5 seconds→ Ask Dr. Math: 5 seconds→ FAQ: 15 seconds → Negative Times Negative: 60 seconds. The user browsed through 5 different nodes and spent different amounts of time on each of the nodes.

*Figure 3. Multidimensional model for **ConQAS***

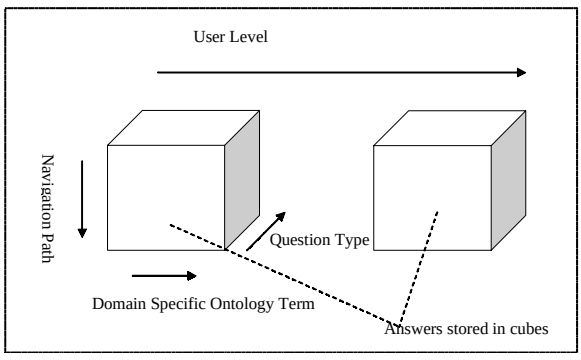
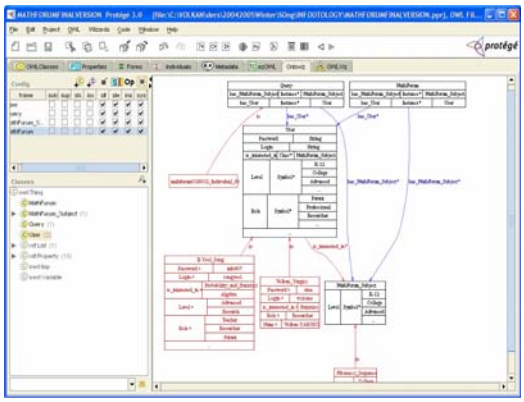


Figure 4. Example of ontology structure



The Second Dimension: Domain Specific Ontology Term.
In our project, we are building the ontology for the MathForum using a bottom up approach and implement it in Protégé 3.0 (http://protege.stanford.edu/). The website of mathforum.org was used as the main information resource. The bottom up approach captures the concepts in the Math Forum, groups them into classes and then defines the associations between the classes to form the ontology with classes and their associations. Higher-level concepts are created by generalizing the low-level concepts and building a hierarchy of concepts using relationships. The following figure is a part of our ontology.

By using the ontology, the question will be preprocessed in order to extract the domain-specific ontology terms. It uses a POS-tagging library and Word Net to identify key words. The selected keywords are checked to see if they appear in the Math Forum Ontology, which is the domain-specific ontology for our target digital library. The extracted ontology terms constitute the third dimension of our multi-dimensional context model.

For example, after a user browsed through the archives as mentioned before, he posts a question as follows. “I have lots of problems with multiplying and dividing integers. Do you have any tricks for me? I can add and subtract integers, but multiplication and division are really hard for me. The negative and positive numbers give me a really hard time.” The keywords identified are “multiplying”, “dividing”, “integers”, “add”, “subtract”, “multiplication”, “division”, “negative”, “positive”, and “numbers”. The system would find out that the ontology terms are “multiply”, “divide”, “add”, “subtract”, and “integer”.

The Third Dimension: Question Type
The question type is used to define the expected answer type, which means the type of information that is being sought. One strong indicator of the question type is provided by the question stem, e.g. when, how, why, etc. Additional resources and mechanisms for identifying the question type are also necessary, because most of the question stems are ambiguous. Vicedo et al. (2001) discuss an additional resource called definition terms. Definition terms are the terms in the question that defines characteristics of the expected answer. For questions with question stems, the noun phrases appearing next to the question stems are considered definition terms. For questions without question stems, the noun phrases following the verb are considered definition terms. Harabagiu et al. (2003) discussed three kinds of additional resources for TREC questions: the dependency structure of the question, mappings from question stems to possible question types that are built in advance, and answer taxonomies that would link the question types to Word Net sub hierarchies.

Questioning our project, we used Vicedo et al.’s definition term to simplify Harabagiu et al.’s (2003) four-step mechanism to find out the question type. The simplified four steps are as follows:

(1) Determine {A}, all the question types corresponding to the question stem.
(2) Determine the definition term N.
(3) Search for the definition term N along all answer hierarchies subsumed by {A}.
(4) Return the answer type as the top of the hierarchy where N was located.

Take the question stem of HOW as an example. All the possible question types corresponding to HOW in our math forum are: NUMERICAL VALUE and MANNER. Similar to HOW, question types corresponding to other question stems would all be built offline and stored in the knowledge base. All these question types would be linked to Word Net sub hierarchies.

When the system tries to find out the question type of a question like “How do you subtract integers like5-(-6)?”, the definition term of this question would be first selected. The definition term of this question is “integer”. The system would then search for this term in Word Net until it is found under the concept of NUMERICAL VALUE, which is the question type of this question.

The Fourth Dimension: User Level
In our current system ontology, there are four different levels of users, which are elementary school, middle school, high school, college and beyond. It’s important to notice that high school students and college students or other advanced mathematics learners expect different answers when they ask the same questions. It is the fundamental difference between different levels of users that make us to include this dimension in our Multi-dimensional context model.

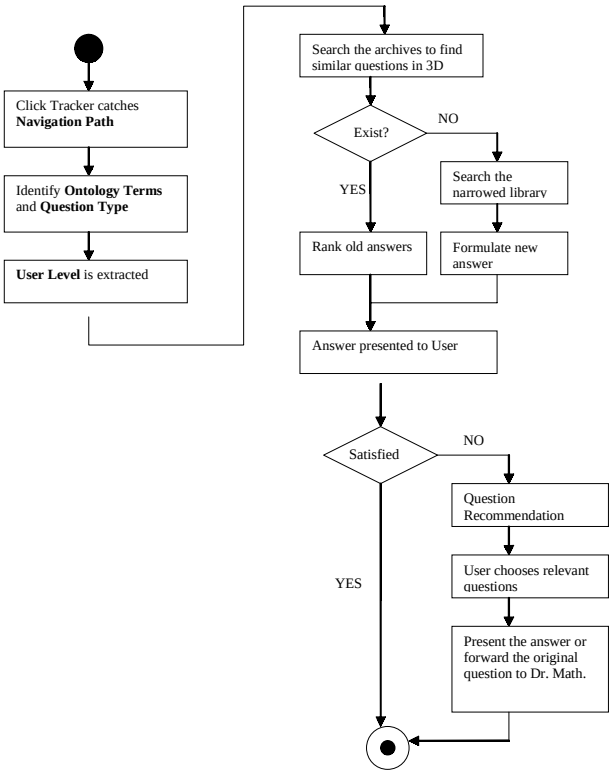
The identification of user level is performed by question analysis function of the Retrieval Manager. The user level is identified by our data mining algorithm from the previously stored data. By using a math dictionary, the system can detect the user level from the key words identified from the question. Consider the above example we used before. The user has the navigation path as earlier mentioned. From the wording of the question like “multiplying and dividing integers”, “negative and positive numbers”, the system can confidently detect that this is a middle school user.

CONTEXTUAL SUPPORT FOR QUESTION ANSWERING

The way **ConQAS** handles questions is different from other Q/A systems. After preprocessing a question, a typical Q/A system would go directly to the digital library and search it in order to find out the relevant document, and then formulate the answer from the retrieved document. For our system, **ConQAS** will follow the following mechanism to find out the final answer.

1. First of all, Click Tracker will capture the navigation path of the user. Then the question will be preprocessed to identify the domain specific ontology term and the question type. User level will be determined using data mining techniques.
2. Search the archives and find out the existing questions similar to the current one. The similarity between questions will be mainly determined by three dimensions (3D): domain specific ontology term, question type, and user level. So, only the questions that are similar to the current one in all of these three dimensions will be found out. If no such questions exist, the system would go directly to Step 5.
3. According to the contextual similarity between the existing questions and the current one, the answers to these questions will be ranked. The contextual similarity here will be determined by Navigation Path. The system has a special algorithm for calculating a path similarity.
4. The answer will be ranked and presented to the user.
5. If there are no reusable answers existing in the archives, the retrieval manager would search the digital library to find out the answer. However, it would not search the entire library. Instead, it would only search nodes that are navigated by user. If it cannot

*Figure 5. High level flow of **ConQAS***



find any relevant documents from these nodes, it would search all the nodes under the certain user level. After a relevant document is found, the answer would be formulated according to the expected answer type.

6. If the user is not satisfied either with the ranked existing answers or the formulated new answer, **ConQAS** would recommend a certain number of questions to the user based on a navigation path similarity.
7. The users would then choose relevant questions from the recommended questions. The system will provide corresponding answers too.
8. If the user still cannot find any relevant questions from the recommendation, the original question will be forwarded to Dr. Math.

The whole process flow of **ConQAS** is illustrated in Figure 5.

CONCLUSION

In this paper, we have presented a framework for the context-aware question answering system (**ConQAS**) for the MathForum Digital Library at Drexel University. We considered four different context dimensions in our system. They are the navigation path, the domain-specific ontology term, the question type, and the user level. The navigation path dimension is defined by two attributes, which are the visited nodes and the time spent on each nodes. The second dimension is the domain specific ontology term, defined by the ontology terms that appear in the question. The third dimension is the question type, which is captured by the question processing function using natural language processing techniques. It requires both information from the question itself, and supportive information from knowledgebase. The last dimension is the user level, which allows us to find different answers even for similar questions.

ConQAS would answer the math questions more accurately by making use of contextual information. The system tries to reuse the archives of Mathforum instead of formulating every answer from scratch. The MD context model enables us not only to reuse existing answers, but also to reuse existing questions.

REFERENCE

Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating contextual information in recommender systems using a multidimensional approach. ACM Transactions on Information Systems (2005) 103-145

Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkerly FrameNet Project. (1997) 86-90

Cheverst, K., Davies, N., Mitchell, K., Friday, A., Efstratiou, C.: Developing a context-aware electronic tourist guide: some issues and experiences. CHI (2002) 17-24

Coutaz, J., Crowledy, J.L., Dobson, S., Garlan, D.: Context is key. Communications of the ACM (2005) 49-53

Dey, A.K., Abowd, G.D.: CybreMinder: A Context-aware system for supporting reminders. HUC (2000) 172-186.

Dey, A.K., Salber, D., Abowd, G.D., Futakawa, M.: An architecture to support context-aware applications. UIST (1999)

Dey, A.K., Abowd, G.D.: A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. http://www.cc.Gatech.edu/fce/context-toolkit

Gross, T., Specht, M.: Awareness in Context-aware Information systems. Mensch & Computer (2001) 173-182

Gyssens, M., Lakshmanan, L.V.S.: A foundation for multi-dimensional databases. Proceedings of the 23rd VLDB conference (1997) 106-115

Harabagiu, S.M., Maiorano, S.J., Pasca, M.A.: Open-domain textual question answering techniques. Natural Language Engineering (2003) 231-267

Hong, J.I., Landay, J.A.: An infrastructure approach to context-aware computing. Human-Computer Interaction (2001) 287-303

Jose Luis, V., Antonio, F.: A semantic approach to Question Answering systems. http://trec.nist.gov/pubs/trec9/papers/alicante_trec_9_paper.pdf

Korkea-aho, M.: Context-aware Applications Survey, http://users.tkk.fi/~mkorkeaa-/doc/context-aware.html

Liddy, E., Diekema, A., Yilmazel, O.: Context-based question-answering evaluation. ACM SIGIR (2004) 508-509

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., Karger, D.R.: The role of context in question answering systems. ACM CHI (2003)

Miller, G.A.: WORDNET: A Lexical Database for English, Communications of the ACM (1995)

Moldovan, D., Pasca, M., Harabagiu, S.,Surdeanu, M.: Performance issues and error analysis in an open-domain question answering system. ACM Transactions on Information Systems (2003) 133-154

Shahabi, C., Zarkesh, A.M., Adibi, J., Shah, V.: Knowledge Discovery from Users Web-Page Navigation. IEEE (1997) 20-29

Small, S., Strzalkowski, T., Janack, T., Liu, T., Ryan, S., Salkin, R., Shimizu, N.: HITIQA: Scenario Based Question Answering. http://acl.ldc.upenn.edu/hlt-naacl2004/qa/pdf/small-hitiqa.pdf (2004)

Stanford University School of Medicine: The Protégé Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/

Surdeanu, M., Moldovan, D., Harabargiu, S.: Performance analysis of a distributed question/answering system. IEEE (2002) 579-596

Vicedo, J.L., Ferrandez, A.: A semantic approach to Question Answering systems.http://trec.nist.gov/pubs/trec9/papers/alicante_trec_9_paper.pdf (2001)

Winograd, T.: Architecture for context. Human-computer interaction (2001) 401-419

Yang, H., Chua, T.: QUALIFIER: Question Answering by Lexical Fabric and External Resources. http://www.comp.nus.edu.sg/~yangh/publication/QUALIFIER-eacl2003-proceeding.pdf (2003)