Kabakuş ve Çetin, Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 30(3):193-197

Erciyes University Journal of the Institute of Science and Technology
Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi
ISSN 1012-2354
Cilt (Volume): 30, Sayı (Issue): 3, Haziran/June-2014
http://fbe.erciyes.edu.tr/

Uzaktan Eğitime Yönelik Anlamsal Soru Cevap Sistemi

A. Talha Kabakuş[1], Aydın Çetin[2]

[1] *Abant İzzet Baysal Üniversitesi, Bilgi İşlem Daire Başkanlığı, 14280, Merkez, BOLU*
[2] *Gazi Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, 06500, Beşevler, ANKARA*

Anahtar Kelimeler: Soru cevap sistemleri, doğal dil işleme, uzaktan eğitim, cevap çıkartımı, e-öğrenme

ÖZET

Uzaktan eğitim ve e-öğrenme sistemleri web teknolojilerinin gelişimi ve artan eğitim ihtiyacı nedeniyle günden güne daha popüler hale gelmektedir. Yazılım çözümlerinin sahip olduğu kararlı yapısından ötürü insan kaynaklı çözümlerden daha kararlı olduğu kabul edilmektedir. Uzaktan eğitim sistemleri ve e-öğrenme portalları web tabanlı iletişimin avantajlarından faydalanarak bilgi edinmede zaman limitini ortadan kaldırmaktadır. E-öğrenme ve uzaktan eğitim sistemlerinin geleneksel eğitime göre en önemli avantajı sorulara cevap alma kolaylığıdır. Bu sistemler zamandan ve mekandan bağımsız olarak öğrenme olanağı sunmaktadır. Mümkün olduğunca çok sayıda soruyu cevaplayabilmek ve kısıtlı insangücü yerine yazılım sistemlerinden faydalanarak DBpedia ontolojisi üzerinden soru cevaplayan bir sistem sunmaktayız. Önerilen yaklaşım esnek bir mimariye sahip olduğundan çeşitli alanlarla ilgili daha özel (spesifik) cevaplar sunabilmek için alanlara özel bilgi bankalarıyla kolaylıkla entegre edilebilir. Geliştirilen sistemde semantik veri ambarları üzerinden sorgulama yapabilmek için doğal dil soruları uygun SPARQL (SPARQL Protocol And RDF Query Language) sorgularına dönüştürülmektedir. Sorulan tüm sorular cevaplarıyla birlikte aynı soruların tekrar sorulması halinde sistemin performansını artırmak için veritabanında saklanmaktadır. Kullanıcı kayıt mekanizması ile öğrenciler sistem üzerindeki soru geçmişlerini sonuçlarıyla beraber görebilmektedir. Uzaktan eğitim ve e-öğrenme sistemleri tarafından kullanılan birçok soru cevap sistemi sadece öntanımlı soruları cevaplayabilmektedir. Geliştirilen ontoloji temelli yaklaşımda ise insangücüne gerek duyulmadan daha fazla sayıda soru cevaplanabilmektedir. Bu yaklaşımla aynı zamanda sistemin geliştirilmesi ve bakımı için gerekli olan insangücü minimize edilmektedir. Bir doğal dil sorusu cevapları aynı olmasına rağmen çok farklı biçimlerde sorulabilmektedir. Geleneksel soru cevap sistemleri bu çeşitliliği ayırt edebilmek için soruların tüm formlarıyla beraber tanımlanmasına ihtiyaç duymaktadır. Fakat geliştirilen sistem semantik ve dilbilimsel çözümlemeler kullandığından ötürü tüm bu çeşitlilikleri kendiliğinden algılayabilmektedir. Sistemin görme engelli öğrenciler tarafından da kullanılabilmesi için sesli arama ve cevap seslendirme seçenekleri de sunulmaktadır. Bunun yanında bu yaklaşım diğer kullanıcılar için soru cevap sürecini kolaylaştırmakta ve hızlandırmaktadır. Sonuçlar ve elde edindiğimiz tecrübeler, geliştirilen sistemin geleneksel soru cevap sistemlerine göre daha performanslı ve daha kolay sürdürülebilir bir sistem olduğunu göstermektedir.

Semantic Question and Answer System for Distance Education

Key Words: QA systems, natural language processing, distance education, answer extraction, e-learning

ABSTRACT

E-learning and distance education are becoming more popular day by day with the development of web technologies and increase of education demands. Software solutions are accepted more stable than human-powered solutions because of their consistent nature. E-learning and distance education portals overcome time limit for information gaining through benefit of web based communication. Key advantage of e-learning and distance education systems than traditional way is easiness of getting answers for questions. These systems offer learning wherever you are and whenever you want. In order to answer as many questions as possible and benefit from software systems instead of limited manpower, we provide a question answering system through DBpedia ontology. Because of our approach has a flexible architecture, it can be easily ported to other knowledge bases in order to offer domain specific answers. We interpret natural language questions into appropriate SPARQL (SPARQL Protocol And RDF Query Language) queries to be able query over semantic data stores. These previously asked questions are stored in database with answers to be used when same questions are asked in order to improve system performance. With user registration mechanism, students are able to see their own question history and their results. Most of question answering systems used by distance education and e-learning portals are only able to answer questions which their answers are defined before by manually. With our ontology based approach, system is able to answer more questions than before without any human effort. This minimizes human efforts for both system development and maintenance. A natural language question can be asked in different forms despite that they all have same answers. Traditional question answering systems need all of these questions to be defined separately in order to sense this variety. Because of benefiting from both semantic and linguistic analysis, the provided system is able to realize these similarities/differences. Also we provide voice search and answer vocalization to make this system can be used by visually-impaired students. Furthermore this approach also accelerates and eases question answering process. Results and experiments show that proposed system offers better performance and maintenance support than traditional question answering systems.

*Sorumlu Yazar (Corresponding author) e-posta:

Kabakuş ve Çetin, Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 30(3):193-197

1. Introduction

Distance education and e-learning are getting more popular with the development of web technologies. Education demand has dramatically increased in last few years. And it is getting harder to offer physical education environment for all students. Universities have student limit for traditional education because of physical incapability. So distance education is becoming more critical than before because it is the only way to overcome this issue. Distance education emerged in response to the need of providing access to those who would otherwise not be able to participate in face-to-face courses [1]. It encompasses those programs that allow the learner and instructor to be physically apart during the learning process and maintain communication in a variety of ways [2]. It has evolved from correspondence schools to delivery mechanisms such as independent study, computer-based instruction, computer-assisted instruction, video courses, videoconferencing, Web-based instruction, and online learning [1]. Advantages of traditional education can be listed as face-to-face communication, togetherness with other students [3] and direct question answering opportunity. Distance education practitioners and researchers have always been concerned with how much interactivity a distance course could provide for students, since interaction is considered a necessary ingredient for a successful learning experience [4]. Education is also evolved through the development of internet and its technologies. Internet offers knowledge sharing and learning environment no matter when and no matter where you are. In order to improve education quality for distance education and meet deficiencies of virtual education environment, students must be able to ask questions and get answers to them as traditional education students do. The Internet and Web technologies provide an answer to the integration of information systems by technical and syntactic levels, while the development of Semantic Web technologies [5] including such as Resource Description Framework (RDF) and Ontologies show promise to handle semantic heterogeneity [6].

Seven principles for implementing new technologies in distance education programs regardless of implementation method can be listed as [7]:

1. Encourage contact between students and faculty.
2. Develop reciprocity and cooperation among students.
3. Use active learning techniques.
4. Give prompt feedback.
5. Emphasize time on task.
6. Communicate high expectations.
7. Respect diverse talents and ways of learning.

Peng et al. [8] present a Case Based Reasoning (CBR) into traditional Q&A system and put forward Q&A engine which uses keywords of the question to trigger the case and sorts the results by the relationship. The proposed engine is able to modify the weights of the keywords dynamically based on the feedbacks of users.

They also extend the 2-layer architecture of CBR to a 3-layer structure to make the system more scalable and maintainable. Liu et al. [9] present a question and answer system that uses Petri Net in order to prevent information conflict that can happen when students ask questions simultaneously while system runs. They introduce Multi-Avatar structure to present the question and answer process and evaluate it with simulation to prove validity of system. Sun and Zheng [10] propose a web answer system based on natural language processing named NL-WAS. They try to identify semantic type of user questions firstly. When the semantic type of question can't be decided easily, system uses both statistical and semantic similarities for computation to select right answers. Ya-Jun and Yun [11] study on a method to construct network education domain ontology and present confirming of question mode based on ontology and the making answers of standardization questions based on transfer network. They discuss semantic similarity model based on knowledge of ontology to solve the making answers of non-standardization questions.

This paper presents our approach to offer natural language question answering system based on semantic web technologies and linguistic analysis. This system is designed to be used by distance education and e-learning services to provide advanced question answering service which is the most critical deficiency of distance education compared to traditional one. It is aimed to make the system available for visually-impaired students by offering voice search and answer vocalization. This paper is organized as follows: section 2 presents architecture of developed system; section 3 presents our methodology to interpret natural language questions into appropriate queries and section 4 presents conclusion and experiments.

2. System Architecture

Students must be able to use question answering system wherever they are and whenever they want. So web-based architecture is the right solution for question answering systems because of its nature. Therefore we developed semantic web-based question answering system. The developed system uses DBpedia[1] as information source to answer student questions. DBpedia allows you to ask sophisticated queries against Wikipedia and to link other datasets on the Web to Wikipedia[2] data [12]. DBpedia is one of the central linked data datasets in Linked Open Data project [13]. According to DBpedia Wiki, DBpedia knowledge base contains more than 3.77 million things, including 764.000 persons, 573.000 places, 112.000 music albums, 72.000 films and 18.000 video games, 192.000 organizations [14]. Users interact with application server(s) through web browsers. All business processes like natural language processing and query preparation and execution are done at server-side.

1 http://dbpedia.org

2 http://wikipedia.org

Application server interacts with both database server(s) and DBpedia endpoint service to find answers to asked questions. All interactions between client and server are handled by AJAX (Asynchronous JavaScript and XML) requests to improve both client and server performance as it is illustrated in Figure 1.

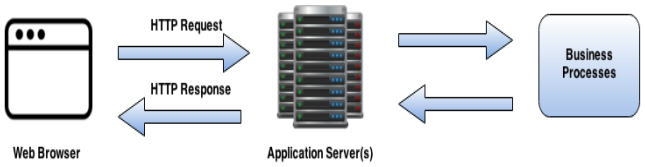

Figure 1. Communication between client and server

System contains a database server for question storage and answer retrieval. System uses DBpedia's query endpoint service to answer questions through DBpedia's knowledge base. We used a NoSQL (Not only SQL) database as storage engine. The reason behind using a NoSQL database (MongoDB[3]) is its better performance for big-data [15] and better scalability for distributed mode. A great example of NoSQL database's performance is Facebook's implementation (Cassandra[4]) which is capable to handle over 100 million users continuously [16]. Another advantage of using NoSQL databases for web applications is that they use same storage format which is JSON (JavaScript Object Notation). So no conversation process is required to render stored data in web user interface.

Figure 2 illustrates system architecture and component interactions. System components with their implementation technologies are listed in Table 1.

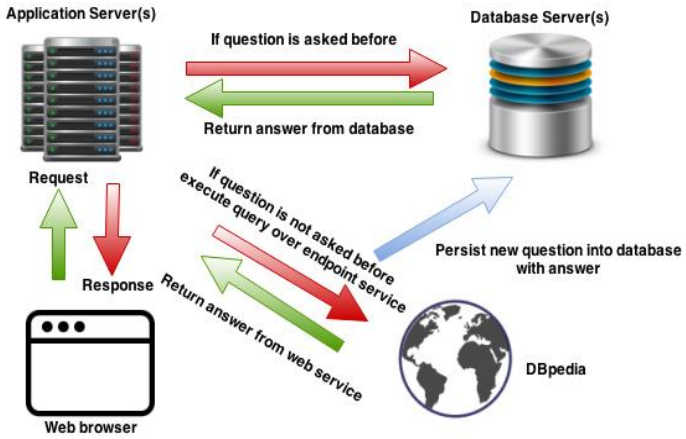

Figure 2. System architecture and component interactions

3. Methodology

DBpedia offers an endpoint service[5] to query its data. In order to use this service to answer questions from DBpedia endpoint service, it is necessary to interpret natural language questions into appropriate SPARQL queries because of service query language limitation which is valid SPARQL query. Our system accepts English natural language questions as input because of its global majority.

Table 1. System components with their

System Component	Technology
Application Server	Oracle Glassfish 3.1.2
Database Server	MongoDB 2.4.7
Semantic Web API	Apache Jena 2.6.4
Natural Language Processing	Apache OpenNLP 1.5.2
Presentation Framework	Sencha Ext JS 4.1
Web Application Framework	Spring MVC 3.2.0
Server Operating System (OS)	Ubuntu 12.04 LTS (Long Term Support)

3 http://www.mongodb.org

4 http://cassandra.apache.org

5 http://dbpedia.org/sparql

Each SPARQL query consists of **|Subject|** **|Predicate|** **|Object|** triple which is similar to English sentence structure. For example we can define "**Turkey's estimated population is 75.627.384**" using SPARQL as "**|Turkey| |populationEstimate| 75627384**". So we need to find equivalent semantic terms of extracted POS tags in order to ask natural question in the form of SPARQL. We used both predefined patterns and DBpedia ontology to construct appropriate SPARQL queries in order to answer as many questions as possible. If the asked question fits a predefined pattern, then SPARQL query is constructed through extracted POS tags. Otherwise SPARQL query is constructed through DBpedia ontology using Levenshtein distance algorithm which is a string metric for measuring differences/similarities between two texts (strings). We determine most similar semantic resources to extracted POS tags through Levenshtein distance values. Question types are determined through these predefined patterns and training datasets that Apache OpenNLP provides. Type checking is used to control whether type of found answer matches with target type or not. Once the appropriate query is constructed, questions are queried over DBpedia using Apache Jena which is an open source Java based semantic web framework offers to execute SPARQL queries on semantic data sources and retrieve results as Java objects. Answers from DBpedia service are retrieved as Java objects and converted into JSON objects to make it ready to present on web user interface. All questions with their answers are stored in database in order to decrease response time once same questions are asked again. System offers optional user registration mechanism to let users to see personalized user interfaces. Otherwise user histories will be limited to their each session. Figure 3 illustrates our methodology to handle natural language questions.

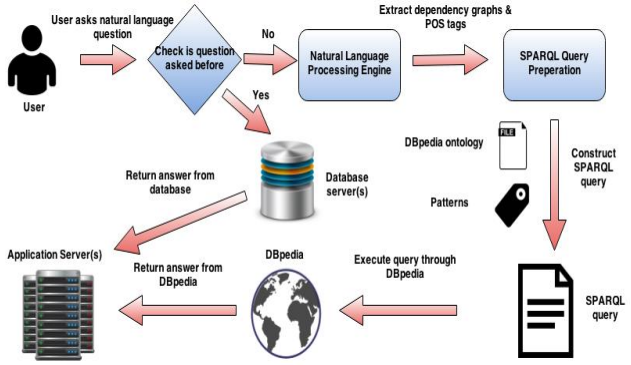

Figure 3. System methodology to handle natural language questions

For visually-impaired students, our system offers voice search and answer vocalization. Students search by voice and their speeches are recognized. As a result of own experiments on speech recognition, we set recognition confidence threshold value to 75% with considering different dialects and hardware deficiencies (deficient sound recorders). This confidence threshold has a range from 0% to 100%. This means that recognition accuracy increases through the rise of this value. When it is set 100%, system becomes too sensitive and doesn't perform well. Table 2 represents system's response rate and recognition accuracy change through this threshold value. If speech is not recognized, an alert is given to warn user.

When speech is recognized as natural language expression, same processes continue. This approach is not useful for only visually-impaired students; it also accelerates question asking process and eliminates spelling errors can happen during question typing. Figure 4 illustrates voice search and answer vocalization process especially developed for visually-impaired students.

Table 2. Recognition accuracy and response change through different thresholds

Threshold	Accuracy	Response Rate
30	Very low	Very high
75	High	High
95	Very high	Very low

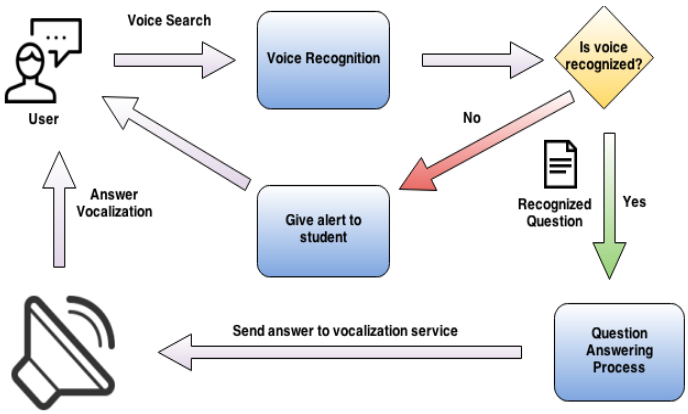

Figure 4. Voice search and answer vocalization process

4. Conclusion and Experiments

We present a semantic web-based natural question answering system in order to answer student questions through a system. Education need of people is growing rapidly despite limited human resource. Traditional education is never enough to demand education need of people. Distance education and e-learning is getting popular day by day because of this limited traditional education service. But they have some deficiencies compared to traditional education like question answering. There are some studies on this area that offers FAQ (Frequently Asked Questions) sections for students. But this approach limits variety of questions that students can ask. Our system doesn't have a predefined question list (FAQ); instead of this we are open to answer any questions can come from students through a well-defined ontology (DBpedia ontology). So we eliminate limits of questions can be answered and benefit from semantic and linguistic approaches to improve both quality and quantity of answers. Our approach can be easily ported to other question domains through changing ontology. We also offer question answering mechanism for visually-impaired users through voice search and answer vocalization which is not offered by any question answering system as we achieved. As a result of considering visually-impaired students, our system is open to more students compared to other works. With user specific web interfaces, students are able to see their previously asked questions with their answer statuses (it is answered or not).

Developed system stores all questions with their answers to accelerate question answering process when same questions are asked. As it can be seen in Table 3, test results and experiments show that this storage mechanism accelerates question answering process by 200-1200%. Because of storing only DBpedia reference values of answers and using a NoSQL database, storage mechanism fairly doesn't affect system performance. This system can be integrated into distance education and e-learning portals to improve interaction capability of their question answering mechanisms.

Table 3. Some results that shows performance improvements gained by caching mechanism

Natural Language Question	**Response Time (sec.)**	**Cache Response Time (sec.)**	**Performance Gain (times)**
Who is Barack Obama?	27	2	12.5
Find places in Germany	28	3	8.3
Who produce Saturnight?	22	2	10
How tall is Michael Jordan?	29	4	6.25
Where is Paris?	20	6	2.33

References

1. Beldarrain, Y.: Distance Education Trends: Integrating new technologies to foster student interaction and collaboration. Distance Educ. 27, 139–153 (2006).
2. Keegan, D.: Foundations of Distance Education. Routledge (1986).
3. Moore, M.G.: Editorial: Three types of interaction. Am. J. Distance Educ. 3, 1–7 (1989).
4. Gunawardena, C., McIsaac, M., Jonassen, D.: Distance education. British Journal of Educational Technology. pp. 403–437 (1996).
5. Shadbolt, N., Hall, W., Berners-Lee, T.: The Semantic Web Revisited. IEEE Intell. Syst. 21, (2006).
6. Thuraisingham, B., Khan, L., Paul, R., Harris, D.: Standards for secure data sharing across organizations. Comput. Stand. Interfaces. 29, 86–96 (2007).
7. Chickering, A.W., Ehrmann, S.C.: Implementing the seven principles: Technology as lever. Technology. 49, 3–6 (1996).
8. Han, P., Shen, R., Yang, F., Yang, Q.: The Application of Case Based Reasoning on Q & A System. AI '02 Proceedings of the 15th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence. pp. 704–713. , Sydney, Australia (2002).
9. Liu, X.L.X., Aoki, Y., Xu, J.X.J.: Development and evaluation of the question and answer system for core courses. Information Technology Based Proceedings of the Fifth International Conference on Higher Education and Training (ITHET 2004) (2004).
10. Sun, X., Zheng, Q.H.: Semantics-based answers selection in question answering system. Advances in Web-Based Learning - ICWL 2004. pp. 354–362 (2004).
11. Ya-Jun, L., Yun, Z.: Question answering system in network education based on ontology and transfer network. J. Harbin Inst. Technol. 36, 1649–1651 (2004).
12. Lehmann, J., Schüppel, J., Auer, S.: Discovering Unknown Connections – the DBpedia Relationship Finder. Proceedings of the 1st SABRE Conference on Social Semantic Web - CSSW '07. pp. 99–110. , Leipzig (2007).
13. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semant. Sci. Serv. Agents World Wide Web. 7, 154–165 (2009).
14. DBpedia Wiki, http://dbpedia.org/About.
15. Leavitt, N.: Will NoSQL Databases Live Up to Their Promise? Computer (Long. Beach. Calif). 43, 12–14 (2010).
16. Okman, L., Gal-Oz, N., Gonen, Y., Gudes, E., Abramov, J.: Security Issues in NoSQL Databases. 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications. pp. 541–547. , Changsha, China (2011).