Dynamic FAQ systems: the state of the art and related work overview

Adam Westerski[1],

[1] Universidad Politecnica de Madrid, Spain
westerski@dit.upm.es

Abstract. In the following paper we present the current state of the art in the area of dynamically generated Frequently Asked Questions systems research. Additionally, since this domain is very little explored, we point of some of the still open problems and discuss related work concepts that are possible to apply in such solutions.

Keywords: FAQ, dynamic, CBR, QA, semantic, web

1 Introduction

In the modern society, access to the right information sources is not only the key element of catering a successful business. Individuals have developed very much the same requirements and needs as companies in the business world. With the growth of the Internet and the so called “Information Society” rapid and precise access to any data has become imminent for almost every person. The availability of information is no longer the only requirement. The delivery speed and accessibility are the key features nowadays. Same as companies, which seek partners who deliver information in the best way, people select products that answer their questions and desires best.

In relation to the contemporary Information Technology domain, the Frequently Asked Questions (FAQ) has been created to satisfy most notably two goals. First: provide users with an easy access to browse the key information about given domain to solve problematic situations. Second: relieve the party responsible for delivering the information of frequently answering the same queries from people interested in the topic.

The FAQ is most commonly perceived as a list of most often asked questions supplied with answers for a given domain. The list is either a selection of most basic questions that the creator anticipates to be often asked or is constructed based on a history of activity in a selected interest group (i.e. users of a selected software application). In the following paper we shall analyze the research done in the area of automatic generation of FAQ lists. Initially, we give an overview of different aspects of creating a FAQ and show the main challenges on the road to automatic content generation (see Sec. 2). Next, we introduce the current state of the art description (see Sec. 3) and follow it with a background of related works in other domains (see Sec. 4). Finally, we conclude the paper with some speculations about the possible future work. (see Sec. 5).

2 Overview

A common Web FAQ published in the Internet environment is most often a static page. The user has to scroll though the entire list to find information that he desires. Depending on the complexity of the FAQ system (see Sec. 3), sometimes it can be supplied with tools for the data management. This includes maintaining a database that stores all the information and lets to easily update the list with new additions and remove the outdated answers. Nevertheless, in practice, most often the FAQ sections, for even very well maintained products, remain unchanged for large periods of time. People responsible for communication with the community tend to acknowledge the need to update FAQ when it is already a critical state and users in very large amount repeatedly ask the same questions.

The dynamic FAQ research is supposed to bring the ability to automatically (or semi -automatically) generate the answers and questions, followed by proper matching. Depending on the scope of the solution a number of problems arise. Some works (see Sec.3) try to solve this problem as one while others decouple it and focus only on partial solutions. Both approaches have their advantages and disadvantages. It has to be noted that the fully automatic solutions are often hyped to be a brilliant use of artificial intelligence but in practice fail to replace manual solutions. On the other hand, simple solutions often still require human input and have similar weaknesses as the contemporary static FAQ systems. Below we present the key aspects of creating a FAQ and point out some possible problems and challenges with respect to introduction of automatic FAQ generation.

2.1 Question extraction or analysis

The first and primary concern for this task is whether or not the questions should be extracted or just entered by the user. Most of the approaches in dynamic FAQ area move towards question-answering systems. In practice, this means the question is formulated by the user. If the input is passed in natural language then the system has to extract the key concepts that describe the question. Next, the extracted concepts should be analyzed and compared with the knowledge base (see Sec. 2.3) to formulate the desired answer (see Sec 2.2). This is a significant change from the original model since the user is no longer presented a list of solutions that he can browse.

The QA systems originally were proposed to support help desk services. Such method is supposed to relive the traffic on call lines and help customers who do not prefer verbal communication. However, in practice those systems are not used often due to unsatisfactory results of natural language processors. In context of dynamic FAQ, an alternative path is to track the available community generated content and try to extract the most frequent discussion topics and the asked questions within. Nevertheless, it has to be noted that this task can prove far more complex and demanding then mere question NLP analysis. So far, the area has not been explored much, however just the initial analysis shows many problems arise. In the ideal situation that conforms to the Linked Data model[1] all community data should be annotated and precisely described. This allows identification of discussion topics and fairly easy question extraction. Nevertheless, in practice the model proposed by the Linked Data initiative is far from being used in practice. Therefore, in order to be applied to existing systems, a solution to question extraction problem would have to take into account natural language processing (NLP) techniques and text analysis methods in order to recognize the actual questions.

Assuming 100% accuracy with question extraction, another significant problem arises. How to determine if a question is asked often or not? It is required to set levels of occurrence frequency that would qualify a certain question to show in the FAQ or not. Additionally, it is obvious that people tend to formulate the same questions using different words and operating on different concepts. Therefore, extracted questions have to be decomposed into concepts just like in QA systems and compared to measure the similarity.

2.2 Answer generation /question answer matching

Once a system has a conceptualized and machine understandable formalization of a question the answer needs to be generated. Once again this problem is valid in more other domains apart of dynamic FAQ or QA systems. It has been vastly explored and depending on knowledge base construction different techniques may be used. The simplest solution is keyword comparison and similarity measure between the question and concepts stored in the database. More sophisticated solutions employ artificial intelligence algorithms or reasoning techniques (i.e. case-based reasoning).

However, it has to be noted that we analyze only a selected domain alternative solutions are possible. For instance, if the questions are extracted from large portions of community data, quite often the questions are provided with actual answers in the same place. The question could be extracted along with the answer and stored in the knowledge base therefore relieving the question- answering process of matching and similarity detection errors.

2.3 Knowledge base

The knowledge base creation is also an important problem in the area. The data used to respond to the questions can be either manually entered or harvested from the available resources. The problem is quite similar as with question extraction. The system can assume it has already an available and formalized knowledge base or it can be designed to analyze the existing natural language information (annotated or unannotated) and formalize it.

3 Solutions

In this section we present various research approaches to building dynamic FAQs. The key differences are in the underlying technologies that those projects use or in the scope with respect to earlier presented dynamic FAQ challenges (see Sec. 2). In the following section we mainly give a general overview of each project and point out its distinctive features or interesting concepts introduced. Some of the technical details of the underlying technologies are described further in Section 4. Additionally, where possible, for each project we point out what exactly is the dynamic aspect.

3.1 Simple Dynamic FAQ techniques

Under this subsection we do not describe a single project but a trend in evolution of static FAQ and the tendency to bring some customization with fairly simple means. There exists a number of contemporary commercial solutions that hype themselves as delivering dynamic FAQ [2,3]. In practice, the idea is to provide a management system with user interface that enables to maintain and modify the FAQ list retained in the underlying database. No complex methods are used but with this simple step a small progress in comparison to static question lists is achieved.

Also sometimes others seek to improve their systems in different ways, that would fit their small scale businesses. For example, a Fiber Optic sales company[4] introduced an improvement to their FAQ based automatic on updates from customer queries sent through emails. Each time a moderator decides a question is worth publishing it is added to the static FAQ list. The manner how this is done is not clear but this simple solution is a good example of real problems that systems operating on the Web face.

3.1 FAQ Finder

The FAQ Finder[5] is a system that puts main focus on question analysis process. A distinctive feature of this solution is that it does not create its own knowledge base. The system uses already existing FAQ files. Moreover, as a result, this research does not investigate answer generation. The user input is analyzed using NLP techniques and compared against the contents of the existing FAQ files. Therefore, the question is not analyzed in order to seek similar concepts in potential answer text but to compare with the list of available precooked questions inside the FAQs. If a similar question is found the corresponding answer is presented to the user.

The techniques to analyze and match the questions are a mixture of NLP and Information Retrieval statistical methods. Usually in the QA systems this technique is used for large data sets that only involve simple question asking (see Sec. 4.1). The solution proposed by FAQ Finder indeed has a big data base of Web FAQs. However usually this type of content contains advice-giving type of questions and answers (i.e "how to ..." or "why...”). The use of Information Retrieval methods for such areas has been proven to deliver bad results and this finds proof in the FAQ Finder experiments that do not bring satisfactory results[5].

3.2 Auto-FAQ

The Auto-FAQ [6] is an implementation of the so called shallow language understanding (see Sec. 4.1). The question/answer matching is based only on the keyword comparison. The Auto-FAQ system in contrast to the FAQ Finder uses its own knowledge base.

3.3 Prioritized Keyword Matching

The solution proposed by Sneiders [7] is an evolution of the shallow language understanding model introduced by Auto-FAQ. In addition to its predecessor it proposes some extensions through the developed Prioritized Keyword Matching mechanism. In relation to the QA systems this is an implementation of template based approach (see Sec. 4.1). Apart of simple keyword matching, three types of keywords are introduced: required (the essence of the sentence), optional (bring additional meaning) and irrelevant (“a” , “the” etc). When two sentences have the same required and optional keywords they are interpreted as a match. Additionally to extend system performance each keyword is perceived as a set of synonyms.

3.4 DynJAQ

The DynJAQ [8, 9] is a very different system then the three previous presented. It proposes solutions for almost each main area of QA process: question analysis, answer generation, knowledge base creation and management.

The questions passed by the user are analyzed and decomposed into basic concepts using NLP. The answer generation algorithm is very dependent on the schema adapted for the knowledge base. The system knowledge is constructed as a directed graph. Each node is a concept that has certain pre-connectors and post-connectors. The pre-connectors determine the required knowledge to learn the concept, while post-connectors reflect the knowledge gained. Therefore an answer to a question is, in practice, a path in the graph that contains a number of concepts extracted during the NLP phase. Such a model of question answering originates from the case-based reasoning. Therefore DynJAQ can be described as a mixture of QA and case-based reasoning techniques.

The authors claim their solution to be adaptable to any domain however in practice they make a great simplification through constructing a FAQ only for JAVA language learning.

3.5 Other tends

Along with the progress of research in the area new attempts seem to follow the trend of applying solutions from other connected domains. The new publications present experiments with techniques such as: semantic matching, vector space model and ranking [10,11], rough set theory and hierarchical clustering [12].

4 Technologies and related research areas

The described dynamic FAQ projects are mostly based on existing technologies or research ideas from other domains that involve text analysis and matching. The innovation in those contributions is based on appliance of various techniques in the particular case of FAQ systems. Although only couple of works have been done to generate FAQs in an automatic way or to introduce some automatic aspects into the process, it can be clearly seen that the authors most often reach out to the question-answering systems or other branches of information retrieval science that fulfill a number of similar needs. Therefore, in this section we try to briefly describe some of the most important research fields that have been already applied in dynamic FAQ. Additionally, as it can be noted further (see Sec 6), the research on Dynamic FAQ does not have limit to those technologies. Along with the new visions of the modern Internet, technologies arise that can be successfully used to dynamically generate user contributed content.

4.1 Question answering systems

The QA research is in a quite mature stage. The first systems regarding this topic have been created in the late 1960ies. Although, since then, numerous new appliances have emerged the ultimate goal has not changed. The question- answering paradigm aims at delivering information in reply to user formulated queries (most often questions stated in the natural language). According to the classification made by Andrenucci et al. [13] there are three main approaches to the topic that emerged throughout the years (see Table 1).

Table 1. Comparison of main QA approaches[13].

	Thorough NLP	IR & NLP	Templates
Entire Web	No	Yes	Yes
Structured Data	Yes	No	Yes
Fact from text	Yes	Yes	No
Advice-giving	Yes	No	Yes
Reliability %	Close to 100	Accuracy > 70	Recall > 80
Small domains	Yes	No	Yes

In reference to heavy use of NLP technique, QA focuses mostly on interfaces to databases [14] and information extraction [15]. The user input is converted into a formal representation (i.e. logic), so that later it can be mapped into a specific database query. The NLP QA systems achieve the best results when applied only to a selected domain, they do no perform well with generic data. In a typical system architecture, the domain- dependent knowledge has a important role during the process of question extraction and database query generation (see Fig.1).

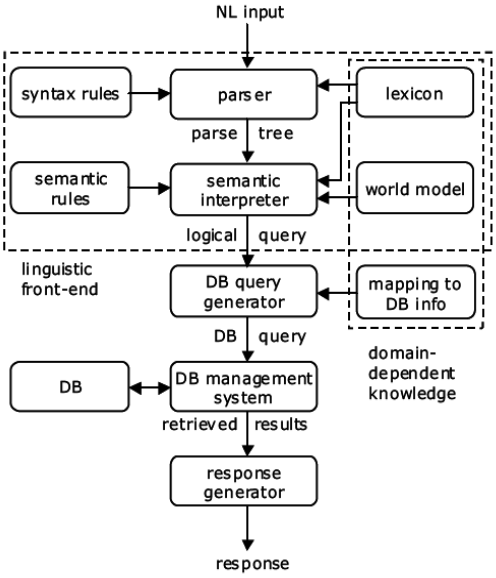

Fig. 1. Architecture of a typical NLP QA system [15].

The second type of QA systems is based on Information Retrieval (IR) techniques. The traditional IR systems are most often employed for document search and retrieval. Therefore, they do not need to be overwhelmingly precise. The user input is matched with the entire document contents- the system does not have to generate a direct answer. With respect to QA, this has evolved into a hybrid solution where IR is used in conjunction with shallow NLP. As a result the precision has increased, while the system is still able to operate on large amounts of data. Under the influence of

QA the Information Retrieval techniques have evolved from document retrieval to passage retrieval[16,17,18]. In general the characteristic feature of those systems is large scale processing and employment of statistical methods that deliver good results when huge amount of data is available. Depending on the amount of NLP enhancements the IR based systems can be language and domain independent.

The third type of QA systems is based on template matching. This technique does not analyze text meaning at all. If a suitable template is available then the relevant information is presented without any guarantee that the answer is correct. In theory the question templates have to be language and domain dependent. Nevertheless in practice many domains overlap therefore basic templates can be often reused. Due to much simpler mechanics then NLP QA this method similarly as IR QA can be used on large scale, although in much more selected cases[19]. In contrast to IR QA, template based QA does achieve quite good results with small subsets as well. The drawback of using templates is lack of any text processing methods therefore facts extraction from text is impossible with this method.

From the point of view of dynamic FAQ the most important feature of the described QA systems is the ability to utilize in advice-giving situations. In practice, this means responding to questions like “how to do...”, “what is...”. Those types of questions most often tend to show up in static FAQ and are asked by people with small amounts of knowledge in a given topic. As it can be seen in Table 1, NLP based and template QA systems perform well in such situations, while IR QA are not suitable (which in the FAQ area has been proved by FAQ Finder[5]).

4.2 Case-based reasoning

The case-based reasoning technique is applied by problem solving systems through gathering the history of previous queries and their answers. When a new problem arrives it is compared to the existing base in search of similar cases. If found the existing cases are reused for the current situation. Additional benefit of such situation is that after solving a problem the CBR systems add it to their knowledge base thus extending their own capabilities in context of future queries. According to the review of CBR systems presented by Aamodt et al. [20] the general CBR processing cycle includes the following steps(see Fig. 2):

1. RETRIEVE the most similar case or cases
2. REUSE the information and knowledge in that case to solve the problem
3. REVISE the proposed solution
4. RETAIN the parts of this experience likely to be useful for future problem solving

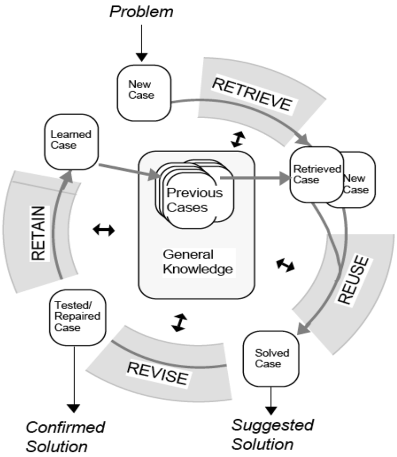

Fig. 2. Architecture of a CBR system[20]

5 Future Work

The basic challenge for all presented dynamic FAQs is the same. Generate answers with perfect accuracy that would ideally match user question. For most of the presented systems, it could be assumed that the task is achieved when they are able to provide correct answer for every question asked. Obviously, this is impossible to achieve but in a large portion the past and contemporary progress in this area is systematically built due to research conducted in related areas (see Sec. 4).

Nevertheless future work in the dynamic FAQs does not have to focus only on text recognition, concept matching and natural language techniques. Like it has been noted earlier (see Sec. 2) dynamic factor can be placed in many other places of the process. For instance, instead of waiting on user input system can maintain and modify the static FAQ list on it own, thus introducing a dynamic element. Additionally, DynJAQ has shown that novel approaches to construction of the knowledge base can provide interesting results. This path could be taken further by introduction of research from other domains such as Semantic Web and RDF stores as a backend for dynamic FAQs. The contemporary trend of Linked Data promotes usage of Web content annotations on mas scale. If applied to community created information (i.e. SIOC initiative[21]) such annotations could possibly deliver valuable data for dynamic FAQs knowledge base.

However, a question arises: where such new paths in dynamic FAQ research could lead ? What could be a potential benefit ? In the end text recognition techniques still need to be used to generate the questions and the answers. I case of Semantic Web and Linked Data(LD), the usage of those technologies could lead to better response accuracy. This is the main benefit that LD brings. With respect to community data published on the Web there is no need to screen scrape or make generic templates for the entire Web. The Linked Data initiative proposes to annotate every single web resource with exact metadata. In practice the drawback of such solutions is the necessity to posses large amounts of such data and all resources actually being annotated. Nevertheless, as the new World Wide Web is being developed, there is a lot of room for experiments.

6 Conclusions

Based on the presented research results it can be seen that task to create good and reliable dynamic FAQ systems that would function without human assistance is hard. The Question Answering systems which are the primary source of technologies for this area are constantly being developed and improved. While they present good results for predetermined and selected narrow domains, the generic engines still perform in an unsatisfactory way. This shows that there is still a lot of room for future research.

Additionally, as proven in the overview and future work sections (see Sec. 2 and Sec. 5) totally different approaches are possible in contrast to the already developed. The scope of problems that QA systems have to solve, and thus dynamic FAQ as well, is very board. The first dynamic FAQ attempts have been initiated more then 15 years ago and still on a wide scale only static FAQs are used. One of the possible reasons for such state might be the overwhelming assumptions and goals of each project. The presented solutions often propose a totally new interaction method for the end user tin contract to the casual FAQ. Additionally each solution requires a lot more input and knowledge from the developers then the contemporary static FAQs construction. It should be noted that the new potential lines of research do not have to deliver a full solution. Small improvements that would only slightly alter the current static FAQ might prove better and more adaptable in the modern Web environment.

References

1. Berners-Lee, T., Linked Data Web architecture note, http://www.w3.org/DesignIssues/LinkedData.html
2. Dynamic FAQ Database, http://products.dynamicwebdevelopers.com
3. Hope Resource Center Dynamic FAQ, http://www.stmarys.org/cancer/faq/default.asp
4. Fiber Optic Store, http://thefiberopticstore.com/FAQ/index.htm
5. Burke, R. D., Hammond, K. J., Kulyukin, V. A., Lytinen, S. L., Tomuro, N., and Schoenberg, S. 1997 Question Answering from Frequently Asked Question Files: Experiences with the FAQ Finder System. Technical Report. UMI Order Number: TR-97-05., University of Chicago
6. Whitehead, S. D., “Auto-FAQ: An experiment in Cyberspace leveraging,” Computer Networks and ISDN Systems, vol. 28(1–2), pp.137–146, 1995
7. Sneiders, E., “Automated FAQ answering: continued experience with shallow language understanding,”AAAI Fall Symposium, pp.97–107, 1999
8. Camacho, D. and R.-Moreno, M. D. 2007. DynJAQ: An adaptive and flexible dynamic FAQ system: Research Articles. Int. J. Intell. Syst. 22, 3 (Mar. 2007), 303-318
9. Camacho, D., “Using Hierarchical Knowledge Structure to Implement Dynamic FAQ System, Proc. Fifth Int’l Conf. Practical Aspects of Knowledge Management (PAKM ’04), 2004
10. Wang, D., Wang, R., Li, Y., Wei B., Latent Semantic Inference for Agriculture FAQ Retrieval, Proceeding of World Academy of Science, Engineering and Technology, Vol. 22, July 2007
11. Wu, Ch., Yeh, J., Lai, Y., Semantic Segment Extraction and Matching for Internet FAQ Retrieval, IEEE Transaction on Knowledge and Data Engineering, Vol. 18, No.7 , July 2008
12. Chiu, D., Chen, P., Pan, Y., Dynamic FAQ Retrieval with Rough Set Theory, International Journal of Computer Science and Network Security, Vol. 7, No. 8, August 2007

13. Andrenucci, A. and Sneiders, E. 2005. Automated Question Answering: Review of the Main Approaches. In Proceedings of the Third international Conference on information Technology and Applications (Icita'05) Volume 2 - Volume 02 (July 04 - 07, 2005). ICITA. IEEE Computer Society, Washington, DC, 514-519
14. Androutsopoulos, I., Ritchie, G. D., Thanisch, P., “Natural Language Interfaces to Databases: An Introduction”, Journal of Natural Language Engineering, 1 (1), 1995
15. Molla, D. et al., “NLP for Answer Extraction in Technical Domains”, Proc. of EACL 2003, Morgan Kaufmann, USA, 2003
16. Salton, G., Allan, J. and Buckley, C., “Approaches to passage retrieval in full text information systems”, in Proc. Of SIGIR«93, ACM Press, N Y, USA, 1993
17. Soubbotin, M. M. and Soubbotin, S.M., “Use of Patterns for Detection of Answer Strings: A Systematic Approach”, in L.P. Buckland and E. Voorhees (eds): Proc. of TREC 2002, NIST, Gaithersburg, USA, 2002
18. Ravichandran, D. and Hovy, EH., “Learning Surface Text Patterns for a Question Answering System”, Proc. of ACL- 2002, ACL press, USA, 2002
19. Lin, J., “The Web as a Resource for Question Answering: Perspective and Challenges”, Proc. of LREC 2002
20. Aamodt, A, Plaza, E., “Case-based reasoning: Foundational issues, methodological variations, and system approaches.” AI Comm Eur J Artif Intell 1994;7:39–59
21. SIOC project homepage, http://sioc-project.org