Knowledge-Based Systems 49 (2013) 81–96

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A cloud of FAQ: A highly-precise FAQ retrieval system for the Web 2.0

M. Romero *, A. Moreo, J.L. Castro

Dep. of Computer Science and Artificial Intelligence, Research Center for Information and Communications Technologies, University of Granada, Spain

ARTICLE INFO

Article history:
Received 9 December 2012
Received in revised form 22 March 2013
Accepted 21 April 2013
Available online 2 May 2013

Keywords:
FAQ retrieval
WordNet
Wikipedia concepts
Natural language
Tag cloud

ABSTRACT

FAQ (Frequency Asked Questions) lists have attracted increasing attention for companies and organizations. There is thus a need for high-precision and fast methods able to manage large FAQ collections. In this context, we present a FAQ retrieval system as part of a FAQ exploiting project. Following the growing trend towards Web 2.0, we aim to provide users with mechanisms to navigate through the domain of knowledge and to facilitate both learning and searching, beyond classic FAQ retrieval algorithms. To this purpose, our system involves two different modules: an efficient and precise FAQ retrieval module and, a tag cloud generation module designed to help users to complete the comprehension of the retrieved information. Empirical results evidence the validity of our approach with respect to a number of state-of-the-art algorithms in terms of the most popular metrics in the field.

1. Introduction

In recent years, Frequently Asked Question (FAQ) lists have received great attention for their capacity to collect and organize user questions and expert answers about specific topics, such as product utilities or practical guidelines. In this regard, a FAQ is expected to be domain-dependant, short and explicit, and frequently asked. FAQ lists are often employed by companies to overcome the costs of technical support departments. New projects capable of exploiting FAQ information are thus of the utmost importance. This research takes the so-called INTER-FAQ project[1] as an starting point, supported by the *Junta de Andalucía*.[2] As main goal, a complete FAQ retrieval framework able to extract useful information from FAQ lists is needed. In addition, the nature of this project imposes two main requirements: (i) the system should deal with large documents collections, and consequently (ii) minimizing the time complexity of the retrieval performance becomes paramount.

As first issue, the framework provide users with a FAQ retrieval process in order to avoid manual searches through the FAQ collection [47]. FAQ retrieval aims to provide users with Question/Answer pairs (from the collection) relevant to users questions expressed in Natural Language (NL). To achieve this goal, choosing a fitted knowledge representation model becomes crucial. On the one hand, statistical approaches are portable and efficient, which is desirable while dealing with large FAQ collections, but they do not explicitly capture the semantic of terms. On the other hand, knowledge-based techniques requires hand-crafted and domain-dependant resources (keywords, linguistic rules, lexicons, domain ontologies, or question templates). Their construction and maintenance process would become a very complex and time-consuming task, depending on the FAQ collection size. Without going further, the Virtual Assistant on the University of Granada web page[3] stored a FAQ list containing over 5000 questions. To manually construct a domain ontology, a lexicon, or a set of keywords (to quote some of them) from this domain becomes a colossal task. Unfortunately, automatic Information Extraction techniques, such as Automatic Keyword Extraction (AKE), Named Entity Recognition (NER) techniques, or Ontology-based annotators do not entirely fulfil with problems above mentioned. Most of them also need additional domain-dependant resources to obtain precise results (the interested reader is referred to [43] for a discussion on Automatic Keyword Extraction focused on FAQ collection).

The second issue of this project involves the Web 2.0. The trend initiated by the peek of Web 2.0 reveals that classical FAQ retrieval systems do not have into account more users' requirements than the retrieved answer[8]. Current systems should provide the user with more information about the query than just a ranked list of pairs. For example, presenting to the user an overview of all the resources available about the query. In this sense, a current Web 2.0 system could be interpreted as a learning resource: it do not represent a type of learning technology, but it can enhance a learning activity notoriously [10,9,25]. We adopt this idea to enhance a

* Corresponding author. Address: C/Periodista Daniel Saucedo Aranda, s/n, E-18071 Granada, Spain. Tel.: +34 958244019; fax: +34 958243317.
E-mail addresses: manudb@decsai.ugr.es (M. Romero), moreo@decsai.ugr.es (A. Moreo), castro@decsai.ugr.es (J.L. Castro).

[1] INTER-FAQ: an information retrieval system based on intelligent FAQs.
[2] http://www.juntadeandalucia.es.
[3] http://tueris.ugr.es/elvira/.

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.04.019

FAQ retrieval system with mechanisms to complete and extend the information requested by the user. Hence, we will define here the concept of FAQ cloud,[4] as an extension of tag clouds [19] following the current tendency in the literature. Tag clouds have become popular in Web 2.0 due to their ability to provide a visual depiction of informative content.

Summarizing, we are responsible for developing an innovative FAQ retrieval system able to (a) manage large volumes of information efficiently, (b) automatically capture the expert knowledge in an interpretable and extendible form, (c) retrieve high-precise answers, and (d) facilitate a domain-learning environment presenting extended information relevant to the user questions. To address these goals, our system implements three main stages. In an initial stage, an information extraction module is responsible for automatically extract weighted information units from the FAQ collection. These units combine the strength of frequency techniques and knowledge-based techniques without needing of human interaction. In a second stage, a query expansion module based on WordNet and Wikipedia prepares the query to the retrieval process. Finally, the retrieval module is in charge of obtaining precise results, and finding and visualizing relevant information. To that end, we design a FAQ retrieval module and a Cloud retrieval module. Effectiveness of our retrieval system is contrasted with state-of-the-art algorithms for FAQ retrieval and tag cloud selection.

The rest of this paper is organized as follows. Section 2 offers an overview of previous work on FAQ Retrieval and tag cloud tasks. Next, we describe the usability scheme of our system. In Section 4, we depict the system architecture and functionality. FAQ retrieval and tag selection algorithms are commented in Sections 5 and Section 6, respectively. The method of analysis and the experimental validation of our modules are outlined in Section 7. Finally, Section 8 concludes with a discussion of results and future research.

2. Related works

In this section, we depict the characteristics of the main approaches of three major topics. We introduce firstly FAQ retrieval approaches. Then, we briefly present tag clouds as common visualization technique for knowledge in a prototypical way.

2.1. FAQ retrieval

In this section, we discuss the main FAQ retrieval approaches related to our work. To that end, we briefly introduce the earlier systems. Later, current approaches are divided into to categories: methods requiring complex knowledge bases and methods that do not. Finally, systems designed to carry out the FAQ maintenance are depicted.

One of the first works in FAQ retrieval task was FAQ Finder [18]. This system employs a NLP strategy involving a syntactic parser to identify nouns a verbs, and performs concept matching using semantic knowledge through WordNet. It uses a vector-space model (VSM) in order to calculate the similarity degree between questions. Another studies of FAQ Finder system in other contexts can be consulted in [24,5,21]. The Auto-FAQ system commented in [53], follows a keyword comparison criterion to implement the question matching in a shallow language understanding perspective. The system in [47] works in a similar way, mixing a shallow language understanding strategy with a keyword comparison technique called Prioritized Keyword Matching strategies. In turn, Ask Jeeves[5] first classifies the FAQ collection into eleven classes, and then performs a keyword comparison. SPIRE [12] is a hybrid CBR and IR system. The system first follows a CBR approach to reduce the number of candidate documents. Then, its INQUERY retrieval engine module processed the documents employing IR techniques. This method present a main handicap: text passages have to be manually labeled. Finally, FallQ system employs a CBR approach [34]. The system represents the domain knowledge by means of manually crafted keywords in order to build Information Entries (IEs). It depends on, therefore, expert knowledge to define the closed domain.

Recent researches could be classified into two categories: those approaches that require much knowledge modeling, and those that do not. (1) The first category normally involves (a) NLP systems and (b) template-based systems. NLP systems aim to obtain a formal representation of NL to give back a concise answer. Template-based systems use of a set of linguistic templates for the matching process. (2) The second category usually involves (c) statistical systems, which match the user queries to FAQ questions by establishing semantic distance measure between them. They are the most usual to deal with large collections. Their main challenge entails defining syntactic links between linguistic structures with the same semantic (i.e., to find which words can be perceived as synonyms in a given context).

Often, the domain knowledge is modeled by domain ontologies.[6] In this line, [60] combines a domain ontology with a probabilistic keywords comparison measure. [59] mixes a template-based approach with a domain ontological model based on keywords with the aim of catching the user's intention. [20] propose an ontology-based system with an assistance module that focus on create a new answer if none of the existing ones is relevant to the query. Another ontology expansion method is proposed in [35]. This system added new manually annotated questions when the obtained similarity score does not exceed the threshold. Next, in [55], an initial classification of the questions into ten question types is performed. The answers in the FAQ collections are then clustered using Latent Semantic Analysis (LSA) and K-means algorithm. The system employs an ontology based on WordNet and HowNet to obtain the semantic representation of the aspects. Finally, the maximum likelihood estimation in a probabilistic mixture model is used as the retrieval process. In [16], the system models the knowledge by means of a domain ontology. In addition, the system includes personalized services based on users' profiles. Apart from the ontology modeling, there are a number of FAQ retrieval systems that make use of a set of linguistic templates to cover the knowledge. The Sneider's template-based systems [47,48] are examples of this kind. They use matching with both regular expressions and keywords in the retrieve process. For further investigation involving knowledge modeling [41,14,54,6] can be consulted.

The main strength of knowledge-based methods is that they provide precise answers in general. However, they imply many knowledge modeling. To overcome this disadvantage, statistical methods have been proposed. These approaches perform without complex knowledge bases. FRACT system [27] performs automatic clustering on a set of previously introduced questions (query logs) to expand them. Then, the system matches the user query not only with the initial set but also with the expanded set of questions. Those query logs are easy to collect and they cover a large language. Other methods follow a hybrid statistical and NLP strategy. As an example we can observe [31]. The NLP module uses a syntactic parser to classify the type of each question. After that, the statistical module takes part performing a comparison between keywords. This work was developed within the international Text REtrieval Conference (TREC) [52] that promotes the design of FAQ retrieval projects. Next, the system in [57] calculates the probabil-

[4] The reader should be aware of the possible confusion that the term 'cloud' could arise w.r.t. the field of Cloud Computing. The present method has nothing to do with it.

[5] http://www.ask.com.

[6] http://www.w3c.org.