A Hybrid Approach for the Management of FAQ Documents in Latin Languages

Christiane Gresse von Wangenheim[1], Andre Bortolon[2], Aldo von Wangenheim[2]

[1]Universidade do Vale do Itajaí, Computer Science
São José, Brazil
gresse@sj.univali.br

[2]Federal University of Santa Catarina - Computer Science/Production Engineering
Florianópolis, Brazil
bortolon@eps.ufsc.br,awangenh@inf.ufsc.br

Abstract. Essential for the success of FAQ systems is their ability to systematically manage knowledge including the intelligent retrieval of useful FAQ documents and the continuous evolution of the knowledge base. Based on our experiences, we propose a hybrid approach for the management of FAQ documents on programming languages written in Portuguese, Spanish or other latin languages. The approach integrates various types of knowledge and provides intelligent mechanisms for knowledge access as well as the continuous evolution and improvement of the FAQ system throughout its life cycle. The principal strength of the approach lies in the integration of techniques from Case-Based Reasoning and Information Retrieval customized to the specific requirements and characteristics of the management of FAQ documents. The approach is currently being implemented and evaluated in the context of an international research project.

Keywords. Text Retrieval, Case-based Reasoning, Natural Language Processing, Knowledge Management
Research paper

1 Introduction

In order to efficiently implement high-quality software systems, programmers require a detailed and broad knowledge wrt. the used programming language(s). However, as the software domain is characterized through rapid technological advances and frequent changes, a challenge is to support the learning of new programming language. In this context, tacit knowledge describing concrete experiences obtained by individuals, for example, through the observation of attempts and mistakes, has been shown to be an important knowledge source that contributes to an effective learning process. A possible form of representing and communicating this type of knowledge are lists of *Frequently Asked Questions* (FAQ), which state a question and its answer provided by a specialist. They explicitly capture know-how and solution strategies as an aid in the search of an adequate solution to a current problem. Thus, they make know-how available on commonly asked questions, which otherwise would be asked again and again to an expert.

In this context, we aim at developing a software system for the management of FAQ documents on questions related to programming languages. Basically, the system has to store FAQ documents in a structured form, retrieve a relevant document for a given question and support the continuous evolution of the knowledge base. Regarding the specific application domain, requirements [5] include the handling of experiential knowledge in form of FAQ documents, extraction of information from textual documents and its mapping into a structured representation of cases and similarity-based mechanisms in order to allow the retrieval of documents with similar but not necessarily identical questions. Furthermore, the system has to be able to deal with queries and documents in natural language. This includes mechanisms for spelling correction, normalization of verbs, nouns and adjectives as well as the automatic extraction of relevant terms. Our approach is developed to fit a whole set of modern latin languages and in our test application the principal language used is Brazilian Portuguese. However, the system has also to be able to handle queries which may be expressed mixed with English jargon of the specific application domain (e.g., "O que é class?"[1]). If, the system is unable to provide a satisfactory answer based on the existing knowledge base, support has to be provided for the manual answer process through an expert. In order to enable the continuous evolution of the knowledge base, the acquisition and integration of new FAQ documents has to be supported. This includes mechanisms for the semi-automatic indexation of new FAQ documents as well as the continuous enhancement and update of general domain knowledge.

Today, many FAQ repositories exist, mainly in newsgroups or in text form as part of a technical manual. However, most of the FAQ repositories do not provide an efficient access to the knowledge contained in the documents. In general, there exist two approaches for the access: visualization and search. Visualization approaches offer an organized collection that can be explored by the user, as e.g., in newsgroups. The problem is that it can be quite time-consuming to find an answer to a specific question. On the other hand, search approaches provide a mechanism that allows the users to explicitly express her/his requirements and to obtain the best results found in the FAQ repository. Traditionally, Information Retrieval (IR) techniques or key-word search are used, as e.g., by Internet search engines. However, these approaches often either overwhelm the user by the amount of retrieved information or are not able to find any useful information. Another problem is the vocabulary problem, which results from the diversity of expertise and backgrounds of the system users. Recently, various systems have been developed which address the problem of handling textual documents by means of knowledge-based techniques, in particular Case-Based Reasoning (CBR) [2]. The principal advantages of CBR in this context are its similarity based retrieval of cases and the incremental evolution of the knowledge base. Specific textual CBR techniques have been developed in order to handle textual cases (e.g., [3, 4, 8, 11]). They allow the integration of any type of general domain knowledge and, thus, allow content-oriented document search strategies which perform much better than traditional IR approaches. However, the existing approaches focus on the handling of documents in English. Today, in this specific application domain, there does not exist a general approach for dealing with queries and documents in latin languages, which are

1.in English: What is a class?

morphologically much more complex than the English language.
Another aspect not further supported, in general, is the continuous evolution of the FAQ repository. An exception is, for example, [1], which supports organizational memory by routing a question, for which the system cannot find an answer, to a human expert. However, the acquisition and integration of the new cases and the appropriate update of the general domain knowledge have to be intertwined in order to comprehensively support the evolution of the knowledge base.
In this paper, we present a hybrid approach that is able to deal with these problems. It integrates techniques from Case-Based Reasoning and Information Retrieval adapted to the specific application domain.

2 Requirements for Intelligent Text Retrieval in Portuguese and Other Modern Latin Languages

A lot of work in the field of content-oriented document search strategies has been done in English and there has been also some isolated work aimed at some other specific languages. There has been performed very little systematic work oriented to the development of more general strategies capable of handling a whole family of languages. Approaches developed for the English Language cannot be generalized, even for other modern anglo-saxon languages based upon analytical grammars like Dutch, since most modern indo-european languages have morphological and syntactic rules that are much more complex than those of English, even if these languages resemble in the general structure. If there is to be developed a FAQ retrieval system that goes beyond simple keyword search, those morphological and syntactic rules have to be taken at least partly into consideration.
There exist a set of modern indo-european languages of the family of the Latin Languages that share syntactic and morphological rules and are candidates for an integrated approach: Portuguese, Spanish[1], Italian, Galician, Catalunian and Reto-Roman. These languages have sufficiently similar syntactic structures that allow the translation from one into another by substituting the words and applying the adequate morphological rules. These languages have the following principal characteristics in common:

1. Syntactic structures are based upon true analytical grammars and do not possess any kind of surface cases[2], not even a degenerated genitive or a possessive. All semantic relations between noun phrases (NP) are defined by the preposition at the beginning of the NP. The rules governing the generation of phrase structures are also very formal and shared[3] between languages.
2. The verbal spectrum of the Latin Language is maintained, with all different verbal modi such as the subjunctive mode and the gerundium mode still in everyday usage. Most verbal modi have a full set of simple verbal tenses, which in the indicative mode can be 6 different simple tenses (1 present, 3 past tenses and 2 futures).

1.Following the international tendency, we call here the Castilian Spanish Language, *Castellano*, simply as “Spanish”, in contrast to Galician Spanish and Catalunian Spanish.
2.For a detailed discussion on surface cases versus deep cases and Linguistic Universals see [9].
3.See [6] for a discussion on syntax rules of natural languages.

Since there are 3 different regular conjugations depending on the infinitive form of the verb and each of the 6 simple tenses conjugates each of the 6 persons differently, there are 108 different morphological rules only for the regular verbs of the indicative mode. The other modi are simpler but similar.

3. Adjectives and articles undergo fully morphological modification in gender and number by the noun or nouns they refer to. Here the terminal symbols are different among languages, but the morphological rules are the same. E.g., in Portuguese the suffix rule for regular adjectives is "o" if the noun referenced is masculine singular, "os" if masculine plural, "a" if feminine and "as" if feminine plural; in Italian the respective terminal symbols are "o", "i", "a" and "e", but the rule is the same. Nouns that can change gender obey a similar rule: e.g., the word "cat" has 4 different forms (in Spanish: *gato, gata, gatos, gatas*).
4. Double negation[1] is not used.

These grammar rule characteristics are, at a first sight, an enormous advantage in natural language processing of these languages since the detailed verbal structures and the morphological interdependencies convey many semantic information that gives context sensitivity to the parsing of phrases and makes discourse analysis easier. However, in order to build a FAQ system, we need something between a simple keyword search and a full discourse analysis. The approach should be able to capture the semantics of key expressions in the question text, and thus must be able to deal with the full morphological variety of these languages, but does not need to perform a full semantic mapping of the NPs of the query.

The approach we developed has been focused on the Portuguese Language as a testbed, but can be applied to any other of the latin languages cited above.

Regarding our application on the management of FAQ documents on programming languages, we can further state the following specific characteristics and requirements regarding the analysis of the queries and FAQ documents:

- useful answers depend on the type of question, e.g., the question "*O que é o controller?*"(What is a controller?) asks for a different type of answer than the question "*Como implementar um controller?*" (How to implement a controller?).
- English jargon is intermixed with the Portuguese question. These are principally names of standard classes or methods (e.g., "*OrderedCollection*" or "*addAll*") or technical standard terms related to object-oriented programming, such as "*object*". The English terms, in general, are not derived or inflected in the queries. However, a specific characteristic of class or method names is that the individual words may be concatenated to one term. Thus, a frequent spelling error is the separation of those terms by inserting blanks between the individual words (e.g., "*Ordered Collection*").
- frequently observed types of spelling errors in the Portuguese terms of queries are:
  - missing character (e.g., "*eviar*" instead of "*enviar*")
  - extra characters (e.g., "*eenviar*" instead of "*enviar*")
  - erroneous character (e.g., "*enfiar*" instead of "*enviar*")
  - pair of transposed characters (e.g., "*envira*" instead of "*enviar*")
  - missing accents (e.g., "*colecao*" instead of "*coleção*")
- normalization regarding the inflection of Portuguese nouns, verbs and adjectives is

1.Like in the French Language: "ne ... pas".

required, in order to identify morphological variants. For example, if plural forms of nouns are not normalized, it may be impossible to find a related case with the respective noun in singular form or vice versa. According to Brazilian Gramatical Nomenclature [Fer86] the normal form is defined as follows:

- nouns: singular, masculine form (i.e., regular plural generation: "*objetos*"⇨"*objeto*" and irregular plural generation: "*imagens*"⇨"*imagem*")
- verbs: infinitive form (i.e., regular forms: "*evita*"⇨"*evitar*", "*evitando*⇨"*evitar*" and irregular forms: "*posso*"⇨"*poder*", "*feito*"⇨"*fazer*")
- adjectives: singular, masculine form (i.e., "*profundas*"⇨"*profundo*")

Normalization is facilitated in our specific application, as queries are always written in present tense and verbs only occur in 1. person singular, 3. person singular or 1. person plural. Prefixes are important for the semantic meaning of the question and therefore may not be separated from the word root (e.g. "*desabilitar*" (*disable*) and "*habilitar*" (*enable*)).

These requirements show that sophisticated support is required for analyzing the natural language queries in order to allow intelligent retrieval and to support the continuous integration of new documents.

3 A Hybrid Approach: The FAQ@System

The objective of the FAQ@System is to provide a tool that, for a given query formulated in Portuguese, finds related FAQ documents stored in a case base in order to help professionals to solve problems and questions arising during the programming of a software system.

As input to the system, the user formulates a question in Portuguese wrt. to the programming language. Relevant terms are extracted automatically from the given query (including the correction of orthographic errors and normalization of verbs, nouns and adjectives). Based on the terms extracted from the given query, the case base, representing FAQ documents, is inquired. Relevant documents are identified based on a set of indexes referencing the content of the FAQ document. By matching the cases with the query using similarity measures, a partial order is induced among the cases of the base. In a first try, the most similar case is suggested to the user. If this case does not satisfactory answer the stated question, the user can request and explore the next ten most similar cases. If the system fails to provide any sufficiently similar case or all cases retrieved do not satisfactory answer the question, the user can request the support of an expert. Therefore, the user's question is stored in the knowledge base (marked as still unresponded). An expert is informed via e-mail about the open question and asked to provide an answer. Once the answer is available, it is forwarded to the user and in combination with the query captured as a new case into the knowledge base. The new case is automatically indexed and mapped into its structured representation. The results of the indexing process are revised by a domain expert, and if necessary, enhanced, for example, by modifying indexes or adding new terms to the general domain knowledge.

The principal aspects of the approach, knowledge representation, natural language text analysis and extraction, similarity based retrieval and continuous evolution are

described in the following sections.

3.1 Knowledge Representation

The information and knowledge in the FAQ documents is represented in form of cases. In order to represent the FAQ documents in an accessible way, the textual description is mapped into a structured representation, which consist of the question text, the answer text, a set of indexes and a question type (see Figure 1).

Case 007	
Question	Como ordenar uma coleção?
Answer	Enviando a mensagem sort para esta coleção
Indexes	ordenar, coleção
Type	modo (tipo 3)

Fig. 1. Example of case representation

The indexes indicate terms of the question text which are relevant for the retrieval of useful cases in the specific application domain enabling an efficient access to the documents. The classification of the cases per type of question expresses the need for different types of answers. Here, the following categories of questions have been identified [5]:

- Definition: questions beginning with "*O que*", e.g., "*O que é uma coleção?*" (What is a collection?).
- Nature or quantity: questions beginning with "*Qual*", e.g., "Quais tipos de mensagens existem?" (Which message types exist?)
- Modal: questions beginning with "*Como*", e.g., "*Como executar um programa?*" (How to execute a program?)
- Utility: questions beginning with "*Para que*", e.g., "*Para que serve o método hash*?" (For what serves the method hash?)
- Example: questions beginning with "*Exemplifique*", e.g., "*Exemplifique a utlização de uma janela.*" (Give an example for the usage of a window)
- Others: questions beginning with any other term, e.g., "*Classes são objetos?*" (Classes are objects?)

Besides the knowledge represented in cases, general domain knowledge is represented in order to provide support for the automatic text extraction, spelling correction and similarity-based retrieval . This includes:

Domain specific vocabulary, which defines indicative expressions for a predictive indexation by restricting normative key-terms that represent terms and common terminology wrt. the specific application domain (e.g., "*classe*", "*executar*", "*rápido*"). The domain specific vocabulary is used in order to automatically extract relevant terms from the query and FAQ documents. Two types of domain specific vocabularies have been separated, one on class names (e.g., "*collection*") and one on commonly used method names (e.g., "*initialize*") wrt. the specific programming language.

Domain specific English-Portuguese dictionary, which contains terms in English used as technical jargon in the specific application domain. In order to enable the automatic translation of these terms into Portuguese, the dictionary represents the English terms and their Portuguese translation (e.g., "*class*" ⇨ "*classe*").

Domain specific thesaurus, which represents relations between domain specific

terms, such as hierarchical relations (e.g., "*OrderedCollection*"⇨"*Collection*"), associative relations (e.g., "*objeto*"⇨"*classe*") and abbreviations (e.g., "*mvc*" ⇨{"*modelo*", "*visão*", "*controlador*"}). The thesaurus enables the consideration of local similarities on index level.

The domain specific vocabularies, dictionary and thesaurus include only terms which are related to the specific application domain.

Normalization rules, which are used for removing or modifying suffixes (e.g., "*...ns*" ⇨"*...m*", as in "*imagens*"⇨"*imagem*"), changes of genus (e.g., "*...oa*"⇨"*...ão*", as in "*leoa*"⇨"*leão*") and conjugation of irregular verbs (e.g., "*faço*" ⇨"*fazer*") in order to identify morphological variants.

General Portuguese vocabulary, which represents a general vocabulary of the Portuguese Language including more than 20.000 terms based on [10]. The vocabulary is used in the spelling correction and normalization process.

Stop list, which includes about 200 words, such as "*aqui*", "*acima*", "*para*"[1], which are very common in the language and, thus, having a low descriptive potential. The stop list is used in the evolution of the domain specific vocabularies in order to exclude domain irrelevant terms.

3.2 Natural Language Text Analysis and Extraction

The query is described by the user by formulating a question in natural language, e.g., "*Como posso ordenar uma Ordered Collection*" (How can I order an Ordered Collection?). The query is automatically analyzed and mapped to an internal representation. The objective of this step is to extract all relevant terms from the question as indexes and to classify the type of question. This includes the following steps:

Tokenization. The tokenization aims at splitting the text into strings of characters delimited by blanks (e.g., {"*como*", "posso", "*ordenar*", "*uma*", "*Ordered*", "*Collection*"}).

Classification of question type. The classification is basically done based on the interrogative pronoun or the adverb being used in the begin of the question. For example, a question starting with "*Como*" is classified as a modal question, asking for the explanation on how to do something.

Extraction of english domain-specific terms. The automatic extraction of english domain-specific terms is enabled through the usage of the domain specific vocabularies on class and method names and the domain specific English-Portuguese dictionary (see Section 3.1). Regarding the specific characteristics of the application domain, the extraction is done by iteratively concatenating subsequent terms (e.g., "*Ordered*" and "*Collection*" to "*OrderedCollection*") and verifying the resulting terms against the domain vocabularies or the dictionary.

Spelling correction. The spelling correction is based on comparing terms with the general portuguese vocabulary. Any near miss with at least 60% similarity is used to substitute the original term in the query. This optimistic strategy has been chosen in order to prevent the omission of any correct term. For example, if the character "*n*" is missing in "*enviar*", another potential candidate could also be the word "*evitar*", with

1.in English: here, above, to

the same degree of similarity to the wrong spelled word "*eviar*".

Normalization. This step aims at the normalization of terms in documents and queries so that morphological variants between the query and a case will match. Through the normalization process each word is converted into its normal form (e.g., "*posso*"⇨"*poder*"). This is accomplished by an iterative process based on a rule-based reduction of words and a verification of the newly created term against the general Portuguese vocabulary. The defined rules (see Section 3.1) basically undo the spelling rules for adding affixes, covering the generation of plurals and other inflections such as verb endings. Exceptions of the rule, e.g., irregular verbs, are listed explicitly.

Extraction of relevant portuguese terms. The extraction of Portuguese terms is done by comparing each corrected and normalized word of the query against the domain specific vocabulary.

Query	
Question text	Como posso ordenar uma Ordered Collection?
Indexes	{poder, ordenar, OrderedCollection}
Type	modo (tipo 3)

Fig. 2. Example of analysis result

For example, the internal representation of the query "*Como posso ordenar uma Ordered Collection?*" as result of the analysis is illustrated in Figure 2.

3.3 Similarity-Based Retrieval

With respect to the indexes and the question type of the query, for all cases in the base a similarity value is computed. This is done by partially matching the indexes and question type using similarity measures on different levels.

Global similarity. The global similarity of a case c_k of the case base wrt. the given query q is calculated by:

$$sim(q, c_k) = \frac{\sum_{(i=1)} simLoc(q_i, c_k) + simType(q, c_k)}{n+1}$$

where $simLoc(q_i, c_k)$ is the local similarity, $simType(q, c_k)$ is the similarity of the question type and n the total number of indexes of the query q. Any case with a similarity above a treshold is considered as a potential answer candidate for the query.

Local similarity. The determination of the similarity between the query and a case is further enhanced through the integration of the domain-specific thesaurus. This allows the consideration of similar, but not necessarily equal index values. The local similarity $simLoc(q_i, c_k)$ of the ith index q_i of the query q and the case c_k is determined by the maximum local similarity value $simLoc_j(q_i, c_{kj})$ of the index q_i with all indexes c_{kj} of the case c_k. $simLoc_j(q_i, c_{kj})$ is calculated by comparing the index q_i to the jth index of the case c_k, considering also the set s_i of similar terms to the index q_i based on the

domain-specific thesaurus:

$$simLoc_j(q_i,c_{kj}) = \begin{cases} 1.0 & \text{if } (\exists(x \in c_{kj})): x = q_i \\ 0.9 & \text{if } ((((\neg\exists(x \in c_{kj}))): x = q_i) \wedge ((\exists(y \in s_i)): y = q_i)) \\ 0 & \text{if } ((((\neg\exists(x \in c_{kj}))): x = q_i) \wedge ((\neg\exists(y \in s_i)): y = q_i)) \end{cases}$$

Type similarity. The similarity of the question types $simType(q, c_k)$ is determined by comparing the question type $type_q$ of the query q and $type_{ck}$ of a case c_k of the case base in accordance to the question types defined in Section 3.1.

$$simType(q,c_k) = \begin{cases} 1.0 & \text{if } type_q = type_{c_k} \\ 0 & \text{if } type_q \neq type_{c_k} \end{cases}$$

Untying similarity. Using the global similarity measure described above a partial order is induced among the cases of the case base. However, various cases can have the same maximum similarity to the given query. For example, given the following situation as illustrated in Figure 3, the global similarity of case1 and case2 is 100%, as

	Query	**Case 1**	**Case 2**
Question	O que é controller?	O que é controller?	O que é model view controller?
Answer		Controller é uma classe ...	Model view controller é um ...
Indexes	{controller}	{controller}	{model, view, controller}
Type	1	1	1

Fig. 3. Example for untying similarity

they both have the index "controller" and are of the same question type as the query. However, as the primary goal of the FAQ system is to retrieve one unique answer, the untying similarity measure is used in order to refine the similarity calculation by a different type of normalization. The resulting $simUntying(q, c_k)$ does not only include the total number of indexes of the query but considers also the total number of indexes of the case:

$$simUntying(q, c_k) = \frac{\sum_{(i=1)}^{..} simLoc(q_i, c_{kj}) + simType(q, c_k)}{\frac{n+m}{2} + 1}$$

where n is the total number of indexes of query q and m the total number of indexes of case c_k.

Continuing the example from above, using the untying similarity measure, case1 is considered as more similar to the query (simUntying(query,case1)=1) than case2 with simUntying(query,case2)=0,67.

3.4 Continuous Evolution of the FAQ@System

In order to continuously improve the FAQ@System and to update the knowledge base,

new cases have to acquired and integrated in the case base. Furthermore, the similarity measure and the general domain knowledge have to be improved and adapted based on feedback from its application in practice.

3.4.1 Acquisition of cases

Each time a query is manually answered by an expert, a new case is acquired, enabling the continuous evolution and actualization of the case base. The new case is based on the question stated by the user and the answer given by the expert. In order to facilitate the integration of the new case into the existing case base, the indexing process is done by automatically extracting relevant terms based on the existing general domain knowledge using techniques as described in Section 3.2. The created case is then revised by the domain expert, who can add, change or delete assigned indexes based on the question text of the case. If, new terms, which are not yet included in the general domain knowledge, become relevant for the description of a case, the domain specific vocabularies, dictionary and thesaurus have to be updated accordingly, as described in the following sections.

3.4.2 Evolution of the domain vocabulary

The domain specific vocabularies have to be updated every time new domain relevant terms become available in order to adapt to changes in the application environment. In order to keep the required manual effort for the identification of new relevant terms as low as possible, new acquired cases are pre-processed in order to point out potential candidate terms. The objective of this pre-processing is to identify candidate terms in a document by determining their descriptive power in the case base. Among the many probabilistic techniques that have been developed, techniques which typically incorporate term frequency and inverse document frequency [12] have been found to be simple and yet very useful [7]. The basic rationales underlying these two measures are that terms which appear more frequently in a document should be assigned higher weights (term frequency) and terms which appear in fewer documents in the whole case base (the more specific terms) should have higher weights (inverse document frequency).
These weights are determined for each term in the question text of the new case which are not yet assigned as index by the following procedure: Using a stop list (see Section 3.1) any non-semantic bearing terms are filtered. The remaining terms are ranked by using the inverse document frequency, where the weight of the term k in document i is represented by:

$$weight_{ik} = tf_{ik} \cdot (\log_2 n - \log_2 df_k + 1)$$

where tf_{ik} is the frequency of term k in document i, df_k the number of documents in which term k occurs and n the total number of cases in the case base. This weight induces a partial order among the terms of the question text, guiding the manual investigation by the domain expert, who finally decides, if a term is added to the domain-specific vocabularies or dictionary.

3.4.3 Evolution of domain specific thesaurus

Every time a new term is added to the domain specific vocabularies, it may also be necessary to update the domain specific thesaurus, if the new term is related to any already existing term. However, the major impediment to the usage of thesauri has been the cost of their manual creation and evolution. Therefore, a pre-processing is done in order to point out potential related terms. Virtually all techniques for automatic thesaurus generation are based on the statistical co-occurrence of word types in text [12], where similarity coefficients are obtained between pairs of distinct terms based on coincidences in term assignments to the documents of the collection.
Thus, each time, a new term *k* is included into the domain specific vocabulary, a term co-occurrence analysis is performed. Therefore, the documents of the case base are represented by a matrix such as shown in Table 1 based on the vector space model:

	T_1	T_2	...	T_k	...	T_m
D_1	tf_{11}	tf_{12}	...	tf_{1k}	...	tf_{1m}
...	...	...	...	...	...	...
D_n	tf_{n1}	tf_{n2}	...	tf_{nk}	...	tf_{nm}

Table 1. Term assignment matrix

where the rows of the matrix represent the individual document vectors and the columns identify the term assignments to the cases. Then, the similarity between the new term *k* and any other term *l* can be measured based on the respective pairs of columns of the matrix. A similarity measure may be defined as [12]:

$$sim(TERM_k,TERM_l) = \frac{\sum_{i=1}^{n} tf_{ik} \cdot tf_{il}}{\sum_{i=1}^{n} tf_{ik}^2 + \sum_{i=1}^{n} tf_{il}^2 - \sum_{i=1}^{n} tf_{ik} \cdot tf_{il}}$$

given term vectors in the form of $TERM_k = (tf_{1k}, ...,)$ where tf_{ik} indicates the frequency of $TERM_k$ in case *i* and assuming *n* cases in the base. As a result a $term_k$-term association vector T_k is computed expressing the similarity of term *k* with every term *l* of the domain specific vocabulary through $sim(TERM_k,TERM_l)$.
All relations with a similarity measure above a treshold are considered as potential candidates and are ordered by their similarity value. This is revised by the expert, who, if appropriate, adds new associations to the domain specific thesaurus.

3.4.4 Continuous improvement through feedback

Due to the fact, that the FAQ document content area and relevant terms may change over time, the continuous tailoring of the domain knowledge and similarity measure for retrieval needs to be supported during the whole life cycle of a FAQ system. This has to be done by carefully analyzing the performance of the system and possible changes in the application context. As a basis for the maintenance, protocols documenting user´s (re-)actions (e.g., the percentage of acceptance of the first answer provided) and

feedback (e.g., on the general experienced usefulness) can be used by the knowledge engineer. Based on a careful analysis of occurring problems the system has to be adapted accordingly. for example, by increasing the global similarity threshold or revising the domain specific vocabulary.

4 Implementation

Based on the presented approach, a prototypical implementation has been developed on the management of FAQ documents on the programming language Smalltalk [5]. The system allows the retrieval of FAQ documents via queries stated in natural language, supports the manual response of questions and provides facilities for enhancing the stored domain knowledge.

The tool is basically a client-server architecture consisting of three logical layers: presentation, application, and data storage. Knowledge, including the FAQ document cases, as well as, domain knowledge (e.g., vocabularies) is stored in a file system. The initial case base contains 200 FAQ documents. The domain-specific vocabulary includes 647 terms related to the programming language Smalltalk and 2.612 class names and 100 method names wrt. a standard Smalltalk image. The domain-specific English-Portuguese dictionary includes 51 terms and the thesaurus 510 terms which have initially been defined through an analysis of the application domain. The application layer provides support tools for the retrieval of FAQ documents, the manual answer process by automatically contacting experts and users via e-mail, the acquisition of new cases, as well as, the enhancement of the domain knowledge. The access to the system is realized through web browsers and e-mail systems via Internet.

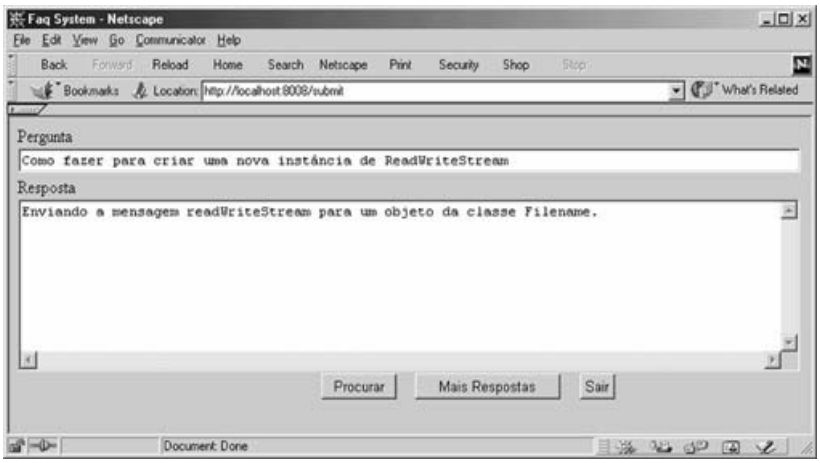

Fig. 4. Example of FAQ@Smalltalk System interface

The tool has been developed platform independent in Smalltalk using VisualWorks 5i.2.

5 Evaluation

We evaluated our approach based on the FAQ@Smalltalk System by adapting the evaluation techniques of CBR and IR systems [12]. Regarding the specific focus of FAQ systems to return one unique answer to the query instead of various potential candidates [4], the following criteria have been evaluated:

- Retrieval speed: the time required for performing the retrieval.
- Recall: the percent of questions for which the system returns a correct answer, if one exists (considering only the first answer provided).

The evaluation has been performed with 40 questions (including orthographic errors, etc.). For 35 questions a most similar case in the case base has been determined by a domain expert as the correct retrieval result. For 5 questions did not exist an adequate answer in the case base.

To evaluate the contribution of the various enhancements made, we performed an ablation study by subsequently eliminating higher level components:

1. complete system
2. without mechanisms for the correction of orthographic errors
3. without mechanisms for the normalization of nouns, verbs and adjectives
4. without mechanisms for the automatic extraction of information
5. without consideration of local similarity
6. without consideration of global similarity (perfect matching only)

The tests have been run using a case base with 200 cases on a Pentium III 800 Mhz with 128 MB RAM.

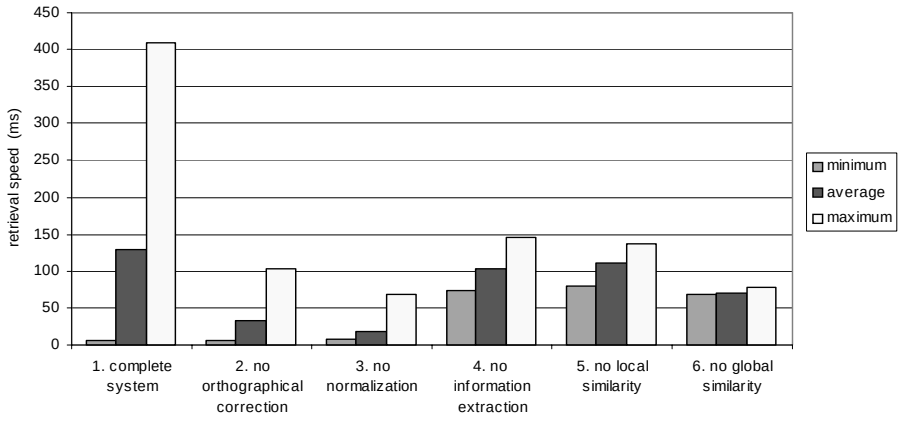

Fig. 5. Retrieval speed

Retrieval speed. In general the retrieval speed is very fast with an average of 129 msec for the complete system. As shown in Figure 5, the speed of retrieval did not increase in accordance with the complexity of the system. For example, we observed a reduction of retrieval time when including information extraction mechanisms. This can be explained through the fact that during the tests 4,5, and 6 more indexes had to be processed (e.g., including irrelevant terms such as articles, pronouns etc.), than during

the tests 1,2, and 3. As a result, the integration of the information extraction mechanisms reduced significantly the retrieval time, resulting in almost the same retrieval time as of a system based on a perfect matching. It also has been shown that depending on the respective need for spelling correction and normalization, the retrieval speed of the complete system can vary significantly from average.

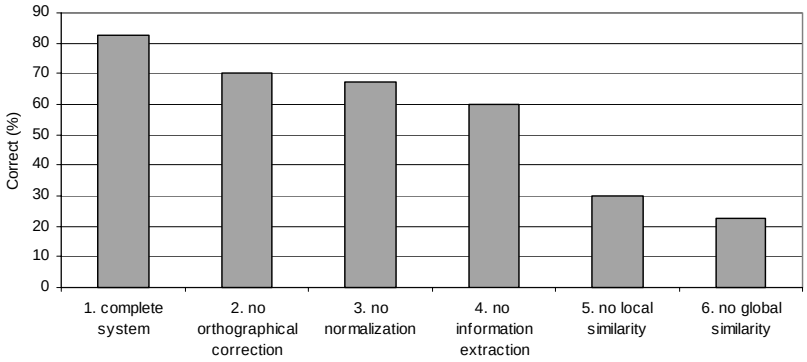

Fig. 6. Recall

Recall. The obtained recall of the approach was high with about 83% of the queries correctly responded by the complete system. 50% of the questions not correctly responded were not covered by the case base. This means that the system returned an answer in a situation where it should not have returned any.
Comparing the different components of the system the largest increase of recall was obtained through the integration of the local similarity measure and the domain specific thesaurus (see Figure 6). Regarding the other components (tests 1,2,3,4) a continuous increase of recall can be observed from about 10% for each new component added.
The evaluation shows that through the integration of various techniques better results regarding retrieval speed and recall can be obtained than by the approaches used individually.

6 Conclusion

In this paper we describe a hybrid approach for the management of FAQ documents in Portuguese by integrating techniques from Case-Based Reasoning and Information Retrieval. In comparison to other existing approaches, the work contributes especially to the automatic information extraction from queries or FAQ documents in Portuguese language and the semi-automatic support for the continuous evolution of domain knowledge. The developed techniques do not only offer an effective and efficient approach for the retrieval of Portuguese documents, but can also be easily adapted to other Latin languages with similar characteristics. The prototypical implementation of the approach is currently being applied in the research group Cyclops at the Federal University of Santa Catarina. Based on feedback from its usage in practice, we intend to direct further research on the amplification of the tool to other FAQ areas (e.g., medical image interpretation) as well as the evolution and generalization of the developed

techniques.

Acknowledgments
The authors want to thank all members of The Cyclops Project who have participated in the initial study and the application of the FAQ@Smalltalk System.

References

[1] M. S. Ackerman. Augmenting the Organizational Memory: A Field Study of Answer Garden. Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW'94),1994.

[2] A. Aamodt, E. Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications, 17(1), 1994.

[3] K. Ashley. Progress in Text-Based Case-Based Reasoning. International Conference on Case-Based Reasoning, Germany, 1999.

[4] R. Burke, K. Hammond, V. Kulyukin, S. Lytinen, N. Tomuro, and S. Schoenberg.. Question Answering from Frequently Asked Question Files. AI Magazine, 18(2), 1997.

[5] A. Bortolon. Desenvolvimento e Implementação de uma Abordagem Hibrida para a Gerência de Documentos FAQ em Portugues. Master Thesis, Production Engineering, Federal University of Santa Catarina, 2001.

[6] N. Chomsky. Aspects of the Theory of Syntax. MIT Press, Cambridge, USA, 1965.

[7] H. Chen, K. J. Lynch, K. Basu, and T. Ng. Generating, integrating, and activating thesauri for concept-based document retrieval. IEEE EXPERT, Special Series on Artificial Intelligence in Text-Based Information Systems, 8(2), April 1993.

[8] J. J. Daniels, E. L. Rissland. What You Saw Is What You Want: Using Cases to Seed Information. In Proceedings of the International Conference on Case-Based Reasoning, Rhode Island,1997.

[9] J. Fillmore. The Case for Case . In Bach, E. and Harms, R.T. (eds.), Universals in Linguistic Theory, Holt, Rinehart & Winston, New York, USA, 1968.

[10] Kuenning, G. H., Karpischek, R. U. International Ispell Version 3.1.20. (ftp://ftp.cs.ucla.edu)

[11] M. Lenz, A. Hübner, M. Kunze. Textual CBR. In M. Lenz et al (eds.), Case-Based Reasoning Technology - From Foundations to Applications. Lecture Notes in Artificial Intelligence 1400. Springer Verlag, 1998.

[12] G. Salton, M. J. McGill. Introduction to Modern Information Retrieval. New York: McGraw Hill, 1983.