APPROACHES FOR CROSS-LANGUAGE INFORMATION RETRIEVAL

[1]Nasir Naveed, [2]Muhammad Tariq Pervez
[1]Department of Computer Science, Virtual University of Pakistan
nasir.naveed@outlook.com
[2]Department of Computer Science, Virtual University of Pakistan
Tariq_cp@hotmail.com

ABSTRACT: *Classical IR system works well for retrieving information from a corpus or internet with having mono language material,l but it is difficult to build a multilingual IR system that can contain documents from the different languages which are used in writing books and content on internet. This involves many issues. In this paper some approaches are devised to deal with different issues and to build a better cross-lingual system.*

Keywords: cross-language information retrieval, CLIR, across languages

INTRODUCTION

The internet contains information and documents in at least 51 languages [1]. Cross-language information retrieval deals with information and knowledge to make it accessible across languages. The first part of the work is the fundamental concepts of information retrieval and the second part expands the concepts of cross-language information retrieval.

1. INFORMATION RETRIEVAL:

Information Retrieval (IR) is a specialized field of computer science, which deals with the computer-based natural language processing. The English term "information retrieval" literally means sourcing information. This requires appropriate representations and possibilities for structuring of complex data stocks and their efficient and content-based [2] search. An IR system tries to satisfy the information needs of a user adding it to a query matching search results.

This part is an insight into the fundamentals of information retrieval. Some essential processes are explained using many detailed examples.

1.1 The naive approach

A user intends to use certain key words to find suitable books; as an example the occurrence of the words "Conspiracy AND Science AND NOT Illuminati". Ensure that all of the relevant books, would it be possible to search the whole library, each book in which the word appears to exclude Illuminati. In the case of a manageable library and suitable hardware this approach will probably work. However, in other areas of applications more efficient procedures and algorithms are required.

1.2 Fundamental strategy of Information Retrieval

Systems are to focus on the actual search request, often referred to as a query is well prepared. In the previous example from 1.1 is the search query "Conspiracy AND Science AND NOT Illuminati". This query will not be processed in one phase. This process can be divided in the three phases that are identified, specific data preprocessing and indexing [3]. The following are the individual phases with examples are explained.

1.2.1 Identification

Depending on the context and the form in which the data to be searched, there may require different procedures for the preprocessing of the data. For example, if a document is in HTML format then data must be extracted from html tags with preprocessing. In order to detect which preprocessing for a specific date is to be applied, this must therefore first be identified. The identification is carried out via the IR system specific heuristics [4], such as the analysis of file extensions or from file headers.

1.2.2 Specific data preprocessing

File identification and preprocessing can make work convenient. After preprocessing a file is converted into a uniform data structure (e.g. simple text). It follows two further preprocessing steps:

(a) The subdivision of this text in tokens
(b) Locate and remove the stop words.

a. The token meaning by disconnecting a text at its spaces. The next step would be to phrases (such as legal person) and proper names such as (Hong Kong, Mercedes Benz, and Lion King) reorganization. Dependencies and relationships of words found by a data-dependence and co-occurrence analysis [5]. For example a co-occurrence analysis shows that in English the tokens Hong Kong almost always occur together. This would, therefore, be a token Hong Kong should be joined together. In the next step the token words are reduced on the basis of its shape and meaning. Several morphological variants of a word in the IR on the basis of this rule are not distinguished. For this activity another technique is used that is called stemming. There are various experimental reasoned algorithms for English language to deliver good results.

b. All obtained tokens come for the indexing contains stop words. As stop words occur very frequently and usually of no relevance for the acquisition to the contents of the document. Often a stop words in the English referred to as an article, conjunctions, prepositions or punctuation. To remove stop words from text, there are lists with appropriate words. These stop words must be matched with words in the list. It is important to note that stop words are only removed after identifying the proper words; otherwise this can distort proper words (King Lion).

Table 1. A token document Matrix

Document Token	Illuminati	Limit	Sacrilege	The Swarm	Pattern Recognition
conspiracy	1	1	1	1	0
Science	1	1	1	1	1
Art	1	0	1	0	0
CERN	1	0	0	0	0
Illuminati	1	0	1	0	0
Space	0	1	0	1	0
tidal wave	0	0	0	1	0

Search query "Conspiracy AND Science AND NOT evaluate Illuminati, logical operations are applied on the Boolean values:

11110 AND 11111 AND NOT 10100
= 11110 AND 11111 AND 01011
=01010

The results of the search for the search query are thus limited and the swarm. At first glance it seems simple and effective, so it is impractical in reality (at least in this scenario). Unfortunately the boundaries of the Boolean IR clear when you look at the example with realistic figures, such as the current books offer of Amazon.com, enriches: for a database with 250,000 books and a total of 100000 vocabulary words (here, it can be assumed that a word a token), provided with a bit of memory per entry is expected to be a table of approximately 2,91GB size will be searched. First of all, this size (for an Amazon Database) is still acceptable, but through this process are not the actual text is found, but only the information in which books the searched words appear. By now also quickly the various texts had to find the documents in much smaller documents, such as only two pages of text, subdivided. The result was that the size of the matrix to a three-digit factor scaled. For books with an average of 400 pages w8*/could indexing table with a size greater than 580GB [5]. By the inverted indexing, one of the most important concepts of information retrieval, the size of the indexing table can be drastically reduced. Around the basic idea behind this technology is to understand it is enough to make the following comment: If on two pages of text in a document (only) a maximum of 1000 words are available, then you are automatically in each vector (one word) 99,000 elements 0. It is therefore possible for at least 99% of the space saving, if only the fields with the value 1 will be stored. This is exactly what happens in the inverted indexing: for each word, which occurs in the database that is stored in the documents it contains. Each document must be clearly identifiable and is therefore with a system-internal document ID. The words are listed alphabetically in the inverted index table is saved so that you do not duplicate may be present.

As previously announced, the three phases of the presented data processing in information retrieval on a big example are demonstrated in Table 2. It will gradually be three different documents in an inverted index table can be entered.

Table 2. Three phases of data processing

Document1	Document2	Document3
<html> <body> Helium Chemical element, discovered by American explorer Bob Moon </body> </html>	A researcher in his Swiss laboratory was found murdered. He was very extremely promoted to soil.	May 2025: The Supply earth seems assured, since America on the moon promotes the element helium-3.
1- Identification of documents and selection of preprocessing		
Document1: HTML	Document2: TEXT	Document3: TEXT
2- Preprocessing: Documents in uniform format		
Helium chemical element, discovered by the American explorer Bob Moon	A researcher in his Swiss laboratory was found murdered. He was very roughly to floor	May 2025: The Supply earth seems assured, since America on the moon promotes the element helium-3

a) preprocessing: token unification and stemming (example)					
Helium chemical element discover by	The American researchers Bob Moon	On researcher becomes in his laboratory murder find	He was very roughly to the floor	May 2025 the supply earth seem secure since	America on the moon promote the element helium 3
(b) remove stop words preprocessing:					
helium chemistry element discover	America researchers Bob Moon	Researchers find murder laboratory	Very roughly soil inquire	May 2025 seem safe supply earth	America promote moon element helium 3

3- Inverter Indexing: Document 1			
Token	Documents in which the Token	Token	Documents in which the Token
America	1	Researchers	1
Bob	1	Discover	1
Chemical	1	Helium	1
Element	1	Moon	1
Inverter Indexing: Document 2 and Document 3			
Token	Documents in which the Token	Token	Documents in which the Token
2025	1	Promote	2,3
America	1,3	Helium	1
Extremely	2	Helium3	3
Bob	1	Laboratory	2
Soil	2	May	3
Chemistry	1	Moon	1,3
Element	1,3	Murder	2
Discover	1	Seem	3
Earth	3	Safe	3
Find	2	Rough	2
Researchers	1,2	Supply	3

This part gave a short introduction to the IR. There were the three preparatory stages; identification, data-specific preprocessing and indexing i presented and explained with examples. It was also the inverted indexing as aimportant concept of the IR is presented and also an example shown. On the basis of these principles, the reader gets introduction of the processes and procedures of the Information Retrieval and lead to the concept of the cross-language information retrieval, to which it should go in the next part.

2 Cross Language Information Retrieval

The difficulty is that when cross-language information retrieval (CLIR) compared to the classical IR is that the language of the search query is different from the documents. In order to meet this difficulty, there are essentially three approaches [7]:

(1) The search request is in real time translated into all languages. Then for each language a separate search request is started.

(2) All documents are should be indexed in all possible languages. Then search queries can be written, translated in the language of classic IR to operate.

(3) All documents and search queries are translated into a main language. This can be a natural language (such as English, Chinese), or an abstract (language-independent) concept of space.

One difficulty that all these approaches have in common is the recognition of the language. So before the various problems of the three approaches are lit up, its first in section 2.1 to identify languages.

2.1 Identify languages

Because when cross-language information retrieval multilingual documents are searched, it is essentially the language of each document once to identify, in order not to risk in English texts polish search queries (or general text on a language to search requests the language B) to process. This step is necessary when indexing of the documents, but at the latest during the processing of the search queries or when you compile the texts.

The different algorithms for language identification in electronic documents are based ultimately on one and the same principle: strings in the text are to be identified with string from a previously trained system compared. This contains information about the frequency distribution of certain strings of all discerning languages. It is obvious that the system included in the trained language with the largest similarity to this text is also the language of the text. The differences between the various voice recognition algorithms are mainly used in the training of the system and the evaluation criteria for similarity of strings.

Within the European Language area for language identification word-based approaches are used. The trained system knows common words and word forms of all languages and their frequency (average frequency of occurrence within a regular text). The training of such a system is relatively complex, since it is half done automatically. Especially for languages with more flexion, the texts with which the system is trained to be very long and are automatically classified as words strings manually checked for accuracy. Nevertheless, this word-based approach to the less cumbersome, on the matching of byte sequences based approach is preferred because (1) for the majority of languages already available, trained systems are available and (2) so the detection of very similar languages is improved.

In order to identify the languages for which no trained system is available or for the word based algorithms are not applicable (e.g. Asian languages), the system is trained with byte sequences instead of words, which is often referred to as N-gram technology. In most cases, this approach is already sufficient. Only languages with very similar byte-N-Programs, which is indicated by the same word strains within similar languages can come about, it, can lead to errors.

They have no difficulty for voice identification Standard documents of a length of more than 20 words, the regular text is included - that is, they contain at least some common function words or other high-frequency word forms. Here the detection rates of all known algorithms over 99% when extremely closely related to the distinction Languages apart.

2. PROBLEMS AND METHODS IN CLIR

In cross language information retrieval, as already mentioned, there are fundamentally different approaches. In the following sections all the individual approaches have been discussed. In addition, problems which they bring with them are discussed [8].

2.2.1 Translate the requirements

The search request is translated in most CLIR systems, to the language of the translated documents to be searched. Then classical information retrieval actions are performed. With the help of dictionary words can be directly translated from one language to another. In the literature in connection with the translation, CLIR often uses multilingual thesauri of speech [8]. A thesaurus is a word network, whose terms by tables are connected to each other. Multilingual thesauri contain tables of equivalence between terms in different languages. With the help of this information words that are semantically equivalent can be summarized. The consequence of this is that without sacrificing quality more relevant search results can be found.

The biggest problem comes when you translate a search request from its length. An average search request has a length of one to a maximum of five words, but already with a length of less than 20 words, is a reliable identification, as mentioned in section 2.

The most words have multiple, partly widely varying meanings. Most of the meanings can in turn by different words are expressed in the target language. This characteristic is called the ambiguity (ambiguity). It is because without context information from the abundance of translation options to select the correct translation.

There are several different methods to disambiguation. Already in part 1 the co-occurrence analysis has been mentioned, the goal of which is to determine how often terms within a contextual framework are mentioned together. With the co-occurrence analysis many techniques can be used with the search query and its context to exclude translations. What is excluded when translation depends on the probability that a system looks for the joint appearance of the searched words. This procedure is problematic if the user precisely needs excluded translations. This can easily happen if the information needs of the user are very special.

A further possibility of disambiguation is because of the grammatical context of a search request to exclude certain meanings. This requires that the search request at least part sets, and this is in German language only in very few cases. In the English language, however this approach makes sense because the same meaning written words often by their grammatical context can be restricted.

It is also distributed the application of so-called query structuring in which all translation variants of terms as synonyms and interpreted in the request through the appropriate operators linked with each other.

To conclude this section should be pointed to two things. There are many different dictionaries and thesauri. The quality of the dictionary is crucial for a good translation of a search query. With the use of so called phrase dictionaries similar result can be achieved by this solution.

2.2.2 Compiling the documents

It is sometimes necessary that all documents should be in all the languages in which the search query may be available. In this case, there must nevertheless be some limitations. Either the number of the searchable documents and of the languages used as far as restricted, that a manual translation of the

documents may be made and makes sense or it will fall back on machine translation with partly strong declines in the quality of the translated texts. For both possibilities there are real scenarios. These translations are done with the intention of improving the information retrieval in the background.

Careful consideration in CLIR approach is required, since the actual IR performs very little cross language functionality, which is required in the previous steps, such as the indexing. For the major part of the language pairs and thus for the practical use the alleged benefits revealed by the translation is not yet sufficiently sophisticated and can be ignored. The machine translation of documents is a very great discipline their development crucial impact on the CLIR.

2.2.3 Search Queries and documents in a uniform language translation

(a) Is there a uniform language to a common language, and then can this approach be used as a hybrid of the two previous approaches under 2.2.1 and 2.2.2. The advantages of this technique are very promising. For example, English as the common language of a system was selected; it would be possible to all IR tools of the English language to use. In addition, the indexing table at a much smaller. Even in difficult-to-process languages such as Finnish, English or Asian languages, it would be easier to the search query to assign the correct result. It remains, however the problems that happen during the automatic compiling error similar to under 2.2.1 and 2.2.2, are so severe that the advantages of little weight. Machine translation in certain languages (1) rapid progress and (2) is easier than the translation into other languages. For example, it is simple because of grammatical structures, a text from the Asian translate into English than vice versa. These selected properties will be used as a common language and could be a particularly favorable approach in future as well.

(b) A completely different method to attempt both documents and queries is to move both in a language-independent concept of space without costly translation processes. This procedure is called Latent Semantic Indexing (LSI) [9]. In LSI the document is search for the concepts. Without individual terms meaning, in such a concept of space a connection between the words Auto, car, truck, car etc. is found and exploited. The procedure is based on the theory that by singular value decomposition of the data the term approximated value frequency and as the authoritative information at the language representation can be used. The consequence of this is that the benefits of the monolingual IR be exploited without on multilingualism to renounce. This procedure fails in practice, however, that such a concept of space so far not reliable and very difficult can be calculated. For the calculation is a multilingual language corpus required. In addition, the mathematical effort of the singular value decomposition is O (n^2*k^3) [10].

3. CONCLUSION

The difficulty of identification of languages can be viewed as a largely resolved problem. For all common languages are sufficiently well functioning heuristics, such as the use of the stop word lists or the n-gram procedure. Since only one of the three presented CLIR approaches, namely the compilation of the search request in all other system languages, at the present time logistically and technically for a large area of application can be implemented and is currently the best procedure. The results that this procedure provides are reusable, but with the linguistic ambiguity because there is no quite sophisticated technology of machine translation to work well. It is depends on the quality of the results of language pair.

The other two methods have their advantages and disadvantages. The compilation of all data in all possible languages can be manually only for a very limited area of use to implement and has its main drawback. But also with progressive technology regarding the automatic translation would be the size of capacity which this procedure takes a disadvantage. An advantage of this technology is the known best quality of the search results. There is a multilingual body; it can be used with the same quality of search results that can be expected with classic monolingual IR. The last method presented has a lot of potential and is based on the resources of the best procedure. It remains to be seen whether the quality of the results with the quality of the other procedures will be comparable and whether more efficient algorithms for transformation into a language-independent concept of space can be found.

REFERENCES

1. Cho, J. and Garcia-Molina, H., "The Evolution of the Web and Implications for an Incremental Crawler", *In Proc. of VLDB '00 Proceedings of the 26th International Conference on Very Large Databases*, 200-209(1999).
2. Bhogal, J., Macfarlane, A. and Smith, P., "A review of Ontology based Query Expansion," *Information Processing and Management Journal*, **43**(4): 866-886(2007).
3. Harshit, S., and Singh, A. K., "A More Discerning and Adaptable Multilingual Transliteration Mechanism for Indian Languages," *In Proc. of 3rd International Joint Conference on Natural Language Processing*(2008).
4. Egozi, O., Evgeniy, G., and Shaul, M., "Concept-based Feature Generation and Selection for Information Retrieval," *In Proceedings of the Twenty-Third Conference on Artificial Intelligence (AAAI)*, **2**: 1132-1137(2008).
5. Schoenhofen, P., Benzcur, A., Biro, I. and Csalogany, K. "Performing cross-language retrieval with Wikipedia," *In Proc. of CLEF*(2007).
6. Argaw, A. A., Asker, L., Coster, R., Karlgren, J. and Sahlgren, M., "Dictionary-based Amharic-French Information Retrieval" *In Proc. of CLEF*(2005).
7. Scannell, K., "The Crúbadán Project: corpus-building for under-resourced languages," *In Proc. WAC- 3: Building and Exploring Web Corpora, Louvain-laNeuve, Belgium,* (2007)
8. Ren, F. and Bracewell, D. B., "Advanced Information Retrieval," *Electronic Notes in Theoretical Computer Science*, **225:** 303-317(2009).
9. Peters, C., Braschler, M. and Clough, P., "Multilingual Information Retrieval - From Research to Practice," *Springer Science and Business Media*(2012).
10. Baker, E. and Rob, U., "Accessing the Compatibility of Document," *Department of Computer Science, University of Sheffield* (2012).