Artificial Intelligence for Information Retrieval

Thomas Mandl
University of Hildesheim, Germany

INTRODUCTION

This article describes the most prominent approaches to apply artificial intelligence technologies to **information retrieval** (IR). **Information retrieval** is a key technology for knowledge management. It deals with the search for information and the representation, storage and organization of knowledge. **Information retrieval** is concerned with search processes in which a user needs to identify a subset of information which is relevant for his information need within a large amount of knowledge. The information seeker formulates a query trying to describe his information need. The query is compared to document representations which were extracted during an **indexing** phase. The representations of documents and queries are typically matched by a similarity function such as the Cosine. The most similar documents are presented to the users who can evaluate the relevance with respect to their problem (Belkin, 2000). The problem to properly represent documents and to match imprecise representations has soon led to the application of techniques developed within Artificial Intelligence to information retrieval.

BACKGROUND

In the early days of computer science, **information retrieval** (IR) and artificial intelligence (AI) developed in parallel. In the 1980s, they started to cooperate and the term intelligent **information retrieval** was coined for AI applications in IR. In the 1990s, **information retrieval** has seen a shift from set based Boolean retrieval models to **ranking** systems like the vector space model and **probabilistic** approaches. These approximate reasoning systems opened the door for more intelligent value added components. The large amount of text documents available in professional databases and on the internet has led to a demand for intelligent methods in text retrieval and to considerable research in this area. The need for better preprocessing to extract more knowledge from data has become an important way to improve systems. Off the shelf approaches promise worse results than systems adapted to users, domain and information needs. Today, most techniques developed in AI have been applied to retrieval systems with more or less success. When data from users is available, systems use often machine learning to optimize their results.

Artificial Intelligence Methods in Information Retrieval

Artificial intelligence methods are employed throughout the standard **information retrieval** process and for novel value added services. The first section gives a brief overview of information retrieval. The subsequent sections are organized along the steps in the retrieval process and give examples for applications.

Information Retrieval

Information retrieval deals with the storage and **representation** of knowledge and the retrieval of information relevant for a specific user problem. The information seeker formulates a query trying to describe his information need. The query is compared to document **representations**. The representations of documents and queries are typically matched by a similarity function such as the Cosine or the Dice coefficient. The most similar documents are presented to the users who can evaluate the relevance with respect to their problem.

Indexing usually consists of the several phases. After word segmentation, stopwords are removed. These common words like articles or prepositions contain little meaning by themselves and are ignored in the document **representation**. Second, word forms are transformed into their basic form, the stem. During the **stemming** phase, e.g. houses would be transformed into house. For the document **representation**, different word forms are usually not necessary. The importance of a word for a document can be different. Some words better describe the content of a document than others.

This weight is determined by the frequency of a stem within the text of a document (Savoy, 2003).

In multimedia retrieval, the context is essential for the selection of a form of query and document **representation**. Different media **representation**s may be matched against each other or transformations may become necessary (e.g. to match terms against pictures or spoken language utterances against documents in written text).

As **information retrieval** needs to deal with vague knowledge, exact processing methods are not appropriate. Vague retrieval models like the **probabilistic** model are more suitable. Within these models, terms are provided with weights corresponding to their importance for a document. These weights mirror different levels of relevance.

The result of current **information retrieval** systems are usually sorted lists of documents where the top results are more likely to be relevant according to the system. In some approaches, the user can judge the documents returned to him and tell the systems which ones are relevant for him. The system then resorts the result set. Documents which contain many of the words present in the relevant documents are ranked higher. This relevance feedback process is known to greatly improve the performance. Relevance feedback is also an interesting application for machine learning. Based on a human decisions, the optimization step can be modeled with several approaches, e.g. with rough sets (Singh & Dey 2005). In Web environments, a click is often interpreted as an implicit positive relevance judgment (Joachims & Radlinski, 2007).

Advanced Representation Models

In order to represent documents in natural language, the content of these documents needs to be analyzed. This is a hard task for computer systems. Robust semantic analysis for large text collections or even multimedia objects has yet to be developed. Therefore, text documents are represented by natural language terms mostly without syntactic or semantic context. This is often referred to as the bag-of-words approach. These keywords or terms can only imperfectly represent an object because their context and relations to other terms are lost.

However, great progress has been made and systems for semantic analysis are getting competitive. Advanced syntactic and semantic parsing for robust processing of mass data has been derived from computational linguistics (Hartrumpf, 2006).

For application and domain specific knowledge, another approach is taken to improve the **representation** of documents. The **representation** scheme is enriched by exploiting knowledge about concepts of the domain (Lin & Demner-Fushman, 2006).

Match Between Query and Document

Once the **representation** has been derived, a crucial aspect of an **information retrieval** system is the similarity calculation between query and document **representation**. Most systems use mathematical similarity functions such as the Cosine. The decision for a specific function is based on heuristics or empirical evaluations. Several approaches use machine learning for long term optimization of the matching between term and document. E.g. one approach applies genetic algorithm to adapt a weighting function to a collection (Almeida et al., 2007).

Neural networks have been applied widely in IR. Several network architectures have been applied for retrieval tasks, most often the so-called spreading activation networks are used. Spreading activation networks are simple Hopfield-style networks, however, they do not use the learning rule of Hopfield networks. They typically consist of two layers representing terms and documents. The weights of connections between the layers are bi-directional and initially set according to the results of the traditional **indexing** and weighting algorithms (Belkin, 2000). The neurons corresponding to the terms of the user's query are activated in the term layer and activation spreads along the weights into the document layer and back. Activation represents relevance or interest and reaches potentially relevant terms and documents. The most highly activated documents are presented to the user as result. A closer look at the models reveals that they very much resemble the traditional vector space model of **Information Retrieval** (Mandl, 2000). It is not until after the second step that associative nature of the spreading activation process leads to results different from a vector space model. The spreading activation networks successfully tested with mass data do not take advantage of this associative property. In some systems the process is halted after only one step from the term layer into the document layer, whereas others make one more step