4th WSEAS Int. Conf. on COMPUTATIONAL INTELLIGENCE, MAN-MACHINE SYSTEMS and CYBERNETICS Miami, Florida, USA, November 17-19, 2005 (pp270-275)

A Machine Learning Approach towards Improving Internet Search with a Question-Answering System

RAINER SPIEGEL
Institute of Medical Psychology
Goethestr. 31/1, Munich D-80336
Ludwig-Maximilians-University (LMU) and Technical University of Munich (TUM) Medical Schools
GERMANY

Department of Experimental Psychology
University of Cambridge (Wolfson College)
Downing Site, Cambridge, CB2 3EB
UNITED KINGDOM

Department of Computing
University of London (Goldsmiths College)
New Cross, London, SE14 6NW
UNITED KINGDOM

Abstract: - This paper introduces a prototype to extract common sense knowledge from the World Wide Web. The prototype combines a search engine with an automated database. It works by extracting information from the enormous amount of documents available on the World Wide Web. Two common examples are that men love women and that women love men (bi-directional relationship) or that boys like toys (unidirectional relationship), whilst toys cannot like boys.

Key-Words: - Machine learning, question-answering system, search engine, automated knowledge extraction, semantic web, human computer interaction

1 Introduction

A recent special issue of *IEEE Computer* addressed the burgeoning role of web-intelligence in an attempt to extract meaningful information from the World Wide Web, e.g. [1]-[4]. In a recent keynote lecture, Lotfi Zadeh presented new challenges to searching the internet more effectively. He focused on the need for question answering systems that are not only able to understand questions in the way they were intended, but also able to provide meaningful answers [5]. Throughout the previous years, a number of question answering systems have been developed. They either address specific or more general purposes. A specific-purpose example is the cross-lingual question answering system for Hindi and English that accepts questions in English, searches for answers in Hindi newspapers and translates these answers back into English [6]. An example for a general-purpose system is a recent development by a group of scientists from Microsoft Research, hence their system is termed AskMSR [7]. General systems such as AskMSR are not context-specific and attempt to find an answer to almost any question. Their strategy is to make use of the World Wide Web as a data repository, because the internet carries a lot of information with a large degree of redundancy. As will be demonstrated later, the approach described in this paper also makes use of redundant online information. A more general overview describing question answering systems linked to machine learning technology can be found in [8]-[9].

Another development that is linked to question-answering systems is the Wikipedia project [10], which gained recent popularity. Wikipedia is an online encyclopedia that differs in 2 major aspects from a usual encyclopedia. First, anyone is able edit the information that appears online. Editing as well as finding an answer works via the World Wide Web, e.g. when trying to find out about the Golden Gate Bridge, one typically finds information that has been entered about the Golden Gate Bridge by other users. Second, the online encyclopedia can be updated continuously. Although Wikipedia is a very powerful project that is being checked for accuracy by authorized editors, one has to bear in mind that it contains several caveats. Users have complained that some of the information provided by Wikipedia is either incomplete or wrong. Because Wikipedia contains a vast amount of information, it is practically impossible to employ enough experts to check every entry in sufficient detail (and several experts on particular topics have

complained that the peer-review process employed by Wikipedia has led to a number of wrong entries). In addition, the fact that volunteers are needed who add/edit manual entries, the system might lack objectivity. Likewise, Wikipedia's speed of growing might be lower than the expansion of an automated system, which requires no interaction with users. Nevertheless, the overwhelming number of volunteers on the World Wide Web might mitigate these caveats in Wikipedia. Since the widespread use of Wikipedia is still relatively recent, the future will have to show how successful it turns out to be. What would be the alternative to a volunteer-based system? As briefly mentioned above, a possible answer is the development of an automated system. Such a system would make use of all available information that can be found for a particular topic when mining the internet. This information could be checked for redundancy. Subsequently, the result could be entered in the encyclopedia database. As will be demonstrated later, the approach I will describe in the following section makes use of this idea.

2 Problem Formulation

The problem with present search engines and question-answering systems is that they have difficulty understanding the full meaning of phrases and that they typically lack a profound understanding of common-sense knowledge. When making a prompt, they are already quite good at choosing relevant websites that refer to the query. However, it is up to the user to skim these websites in order to find the correct answer to the question. Unfortunately there are many cases where a meaningful answer is not even provided. In contrast, the advantage of a system that understands common-sense knowledge is that it would enable a more natural dialogue between users and search engines. Such a system would probably be better able to understand the intention behind a question. The following example is one way of building such a system: permitting users to provide common sense knowledge by entering phrases such as "men love women", "women love men", etc. Due to a potentially high number of volunteers all around the world, such a system might be fed with a large amount of information (e.g. the Wikipedia example). Once a large amount of common sense knowledge has been entered into the system, it could be used to enhance the clarity of queries. The problem is that the data entries would have to be checked before such a system could support search engines. Otherwise there would be no control of the type of information that users enter, e.g. users with malicious intentions could deliberately enter a large amount of wrong information.

An alternative way of creating a system with common sense knowledge is by automating the underlying process. This is the approach taken in this manuscript. An algorithm is proposed that extracts knowledge from the large amount of redundant information that can be found on the World Wide Web. This information appears in documents that have been created in a variety of formats, such as html, doc, pdf, xml, asp, cfm etc. For the purpose of the algorithm, the format is irrelevant. The important criterion is the type of knowledge that is communicated through these documents, e.g. an html document might contain the phrase "boys like toys." The same phrase might be contained in a pdf document. What matters is the frequency with which a particular phrase can be found on the internet, e.g. the uni-directional relationship "boys like toys" is much more frequent than the reciprocal phrase "toys like boys." In bi-directional relationships, reciprocal phrases such as "women love men" / "men love women" have similar frequencies to each other. As will be demonstrated in the next section, these frequencies can support the development of a database that contains a large amount of common sense knowledge. The ultimate aim is to use this system in order to develop search engines that understand users' queries better due to the knowledge having been fed into them. As a result, these search engines might provide better answers to queries.

3 Problem Solution

Common sense knowledge is contained in many documents on the internet. Consequently, its frequency is high. A high frequency of a particular phrase can be detected with classical search engine technology by entering a phrase in quotation marks, e.g. when entering "the rat eats the cheese" into a search engine like Google, it will search for this phrase in this particular word order and display the exact number of documents containing this phrase. Instead of entering the phrase manually, servers such as the ones used by classical search engines could automatically crawl through the excessive number of documents and determine the frequency for every phrase found. Once a phrase is found, it becomes a candidate for a phrase representing common sense knowledge. This phrase could then be modified to see whether other combinations of word order exist. This works by storing each word in a character array and changing word order as well as adding plural forms to the particular words (the respective plural forms can be represented in an online library to which the server has access). By doing this, only the two phrases "the rat eats the cheese" and "rats eat cheese" would be frequent. No hits would be found for phrases such as "the cheese eats rats" or phrases that violate grammar. Occasionally, paradoxical phrases can be found as well, e.g. the phrase "the cheese eats the rat" produced 2 hits. How these occasional examples are dealt with will become clearer when referring to the mathematical part of this paper. Frequent or less frequent

non-contradictory phrases can be stored in a database. Once the database grows, it will represent a question-answering system for common sense knowledge. When asking the system what rats eat, it would output this phrase among other examples such as rats eat crop etc. Having a system represent common-sense knowledge can in turn be applied to improving search. A search engine that is built on the common sense knowledge of its users will be better able to fulfill the users' requests by simply understanding their queries better, i.e. the search engine would be linked to the relational database management system (e.g. via Posgres, mySQL, OracleSQL) where the common sense knowledge is stored. Subsequently, incoming queries are processed to select appropriate answers. Once a query such as "what do rats eat" is sent to the search engine, it would link to the database and output all relevant details of what rats eat, e.g. cheese, crop, plums, garbage etc. To demonstrate the advantage over classical search engines, it has to be kept in mind that these classical systems provided all the URLs of websites that broadly related to this question. To name an example, Google produced over 2 million hits, where most websites broadly refer to this question by naming one or two things that rats eat. Looking through all of the 2 million hits would be impossible, nor would the user have any idea about the type of food rats are connected to most. In contrast, the system presented in this paper can link to the database management system and answer the question in the way it was intended, by naming the things that rats eat. The type of food that had the highest frequency on the web would appear on top of the list and food that had a lower frequency would appear below that. Consequently, this approach might be perceived to be more user-friendly. This is because the user gets a list of things what rats eat rather than having to look through several websites individually to be left with only a rudimentary impression.

3.1 Mathematical foundations

The system presented in this paper makes use of probabilities to a large extent. However, probabilities alone would be insufficient, as demonstrated in the following example: The phrase "men love women" displayed a total of 22,900 hits, whilst the phrase "women love men" displayed a total of 18,200 hits in Google. Does this imply that the probability that men love women (55.7 %) is larger than the probability that women love men (44.3 %)? This is certainly not the case. When implementing the raw probabilities into the system, the bi-directional link would not represent common sense knowledge at all, e.g. it is not sensible to assume that there is a roughly 56 percent probability that men love women and a roughly 44 percent probability that women love men. In reality, there is typically a strong bi-directional relationship for both facts. The question is how an intelligent algorithm can capture this common sense knowledge in an accurate way. For this purpose, equations were applied that have already been used in the context of fuzzy rules. These equations have originally been developed to understand human learning and memory in psychology experiments [11]-[13].

In the approach taken here, probabilities are the starting point. The fact that the percentage of the phrase "men love women" is similar to the percentage of its reciprocal "women love men" gives rise to a bi-directional relationship. Due to the similarity in percentage, both phrases should have an approximately equal chance of being activated in the following equation, where activation depends on a non-linear process and changes dynamically. Excitation increases the activity of a particular phrase, whilst the activity is decreased through decay. In the following Equation (1), excitation and decay appear separately. A certain phrase such as "women love men" is activated with a certain probability p (in this case p=0.443) and is not activated with 1-p (in this case p=0.557). For the reciprocal "men love women", probabilities are the other way round. When running this algorithm, either of the 2 phrases will be activated. Whenever a phrase is not activated, its activity decays (in this case this is the reciprocal to the one that is activated). The decay is dependent on the number of previous activations of this particular phrase (the more often a phrase has been activated in the past, the slower its rate of forgetting). In addition, there is a real-numbered parameter ς that can adaptively change between applications. This value will be referred to at a stage when the reader is more familiar with the equation. The number of previous excitations of a specific phrase is denoted by η.

$$a_{i+1} = 1 - \left[\frac{1}{1+\beta\left(e^{\left(\left(\frac{\beta}{1-a_i} \right)^{\frac{1}{2}} \right)} \right)} \right] + \lambda \left[a_i - \frac{a_i}{1+\eta\left(\frac{1}{\varsigma} \right)} \right] \quad (1)$$

When a phrase is excited (in this case β=1 and λ=0), the excitatory part of the equation becomes active and the decay part of the equation remains inactive. The excitatory part of the equation is the part left of λ, whilst the decay part is right of λ including λ itself. During decay, the decay part will be active (in this case λ=1 and β=0). In a more formal way, the relationship between λ and β can be expressed as λ=(1-β) and β {0, 1}. The activity itself is denoted as α_i and the following activity

(resulting from either excitation or decay) becomes α_{i+1}. If α_i reaches a value of 1, it will be corrected to 0.999, as the function would be undefined otherwise. In this equation, *e* stands for Exponential (Euler's number 2.718) and ensures that the activity is constrained to boundaries between 0 and 1. Because the equation combines several aspects, it may look complex at first sight. It becomes more readable, however, if one considers that only excitatory *or* decay part of the equation can be active at any one time.

At this point one might ask how a phrase such as "women love men" is trained by applying the algorithm that makes use of Equation (1). Before training can take place, the total number of training trials needs to be calculated. This happens in the following way: there were 18,200 search results (44.3 %) for the phrase "women love men" and 22,900 search results (55.7 %) for the phrase "men love women." This is a total of 41,100 hits. Consequently, the number of training trials will be set to 41,100 in this particular example. Each single trial out of the total of 41,100 trials, the algorithm makes sure that there is a probability of exciting the phrase "women love men" with a 44.3 percent probability. The probability that the activity of this phrase decays is therefore 55.7 percent (100 % - 44.3 %). For the reciprocal phrase "men love women", the probabilities are the reverse. After 41,100 trials have been reached, both phrases have reached certain activities. It is important to keep in mind the following: the fact that decay is more likely than excitation for the phrase "women love men" does not mean that this phrase is likely to end up with zero activity. This is due to the non-linear nature of the activation function. Take the following example: after the first excitation, activity will be 0.73 (this follows from entering the values $\beta=1$, $\lambda=0$ and $\alpha_i=0$ into the equation). Although the phrase "women love men" is actually more likely to decay than to be excited at the next trial, it will on average be excited slightly less than every 2nd trial (because this phrase has a probability of slightly less than 50 percent). At the stage of its next excitation, its activity has probably not decayed to zero yet. The next excitation thus adds to the previous activity. In addition, every excitation of this phrase will lead to slower decay, as decay in Equation (1) depends on the number of previous excitations. This is analogous to human memory [11], where people are less likely to forget well-established memory traces (e.g. due to many episodes of learning, there will be less decay and the memory trace remains strong). Whilst the activity of the phrase will build up quickly by reaching a value of 0.73 after the first excitation, it will also decay quickly if this phrase is not refreshed. If this phrase has been repeated many times, however, decay will be much less. The power of ½ in the excitatory part of the equation is to make sure that activity does not build up too rapidly, hence smoothing the learning curve (depicted in Figure 1).

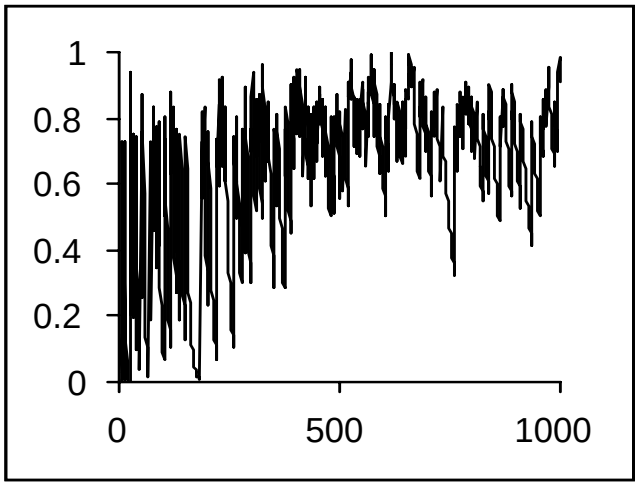

Figure 1: Activity is displayed on the Y-axis, whilst the number of training trials appear on the X-axis [12]. The curve goes up during excitation and down during decay. Decay becomes less steep as the number of previous excitations increases.

When testing the prototype, both phrases "men love women" and "women love men" reached an equally strong activation of 0.99. This indicates that there is a strong reciprocal relationship in both directions. It is common that men love women and vice versa. It is important to note that same-sex relationships are also represented on the internet, thus there are strong activities for "men love men" and for "women love women" as well. All these phrases are very frequent on the internet. Therefore, the previously mentioned decay parameter ς is robust to changes. When trying to extract knowledge from a limited number of phrases (with longer phrases being more wordy and thus less likely to appear in exactly this word constellation), we might still want to reach high activities. This can be done by lowering the value of the decay parameter ς as the number of words per phrase increases.

In the next example, this prototype was tested on a phrase where no reciprocal exists. It is likely that "boys like toys", but impossible that "toys like boys", though this phrase is certainly possible in another context such as "…playing with their transformer *toys like boys* anywhere…" When the prototype of this system was tested, there were 173 "boys like toys" phrases compared with only 3 "toys like boys" examples on the web. Consequently, the total sum of training trials was 176, with a 98.3 percent probability of exciting the phrase "boys like toys" and a 1.7 percent probability of exciting the phrase "toys like boys." When testing the prototype, this resulted in an activity of 0.98 for the phrase "boys like toys" and 0.001 for the phrase "toys like boys."

There is a low chance that on trial 176, the opposite phrase "toys like boys" is excited and thus reaches a relatively high activity of 0.73. Such a high activity would clearly contradict the low frequency of the phrase "toys like boys." To prevent this type of recency-effect, averaging over the past 20 percent of activities provides a more realistic result. When searching for this phrase on the web, there have been changes in the number of hits between the time when the prototype was tested and the present day, but the proportions stay similar, i.e. the system's finding of a strong uni-directional relationship for "boys like toys" persists.

One might ask why such an algorithm with strong effects of excitation or decay is needed at all. What about an algorithm where excitation and decay occurs gradually? This important question becomes more obvious when one considers that there are a number of rare relationships on the web, e.g. the German phrase "München ist Landeshauptstadt des Freistaates Bayern" (Munich is the capital of the Free State of Bavaria) produced a total number of 3 hits in Google, but no other city produced any hits (e.g. when entering "ist Landeshauptstadt des Freistaates Bayern", no alternative cities appeared). By having 3 consecutive excitations throughout the 3 training trials (there is no decay, as there was no contradictory phrase on the web), it is possible to get a high activity (approx. 0.95) for this fact. This would have been impossible with gradual training. If there is both a low number of hits as well as contradictory information, it is less clear what the true information is, so the contradictory phrases should not be implemented with high activities into the database. Hence, there should be significant inhibition especially when there is a low number of training trials. This is made possible through strong decay at the start of training which only gets weaker after a significant number of previous excitations due to recurring consistent (=non-contradictory) information have occured. This aspect has been considered in Equation (1).

3.2 System flow-chart and summary

Testing the prototype of this approach revealed that it is possible to extract a potentially large number of common-sense knowledge from the internet. Storing this information can help the user find a clear answer when sending a query, e.g. the prototype gives the correct answer when being asked about the world's largest city (Mexico City). This not only works in English, but also in other languages (e.g. Spanish). Consequently, the purpose of this paper is to demonstrate the advantage of extracting information from the World Wide Web when aiming to enhance the power of search engines. Although a fully functional prototype is encouraging, this project requires extensive resources and might take time until completion. Nevertheless, the prototype results obtained so far have advantages over many present question answering systems. Systems such as AnswerBus [14] provide URLs that broadly relate to the question's wording, but often do not answer the question as they lack a relation to word order, e.g. some of these URLs contained answers such as *the City of Collinsville (Illinois) has the world's largest water tower* (after having been prompted about the world's largest city). The advantage of the prototype presented in this paper is that its reference to word order makes answers of this type impossible. Rather, it favors candidates that are represented on the World Wide Web in exactly this word order, among them Mexico City (which is in fact the world's largest city) and Tokyo (the 2nd largest city in the world).

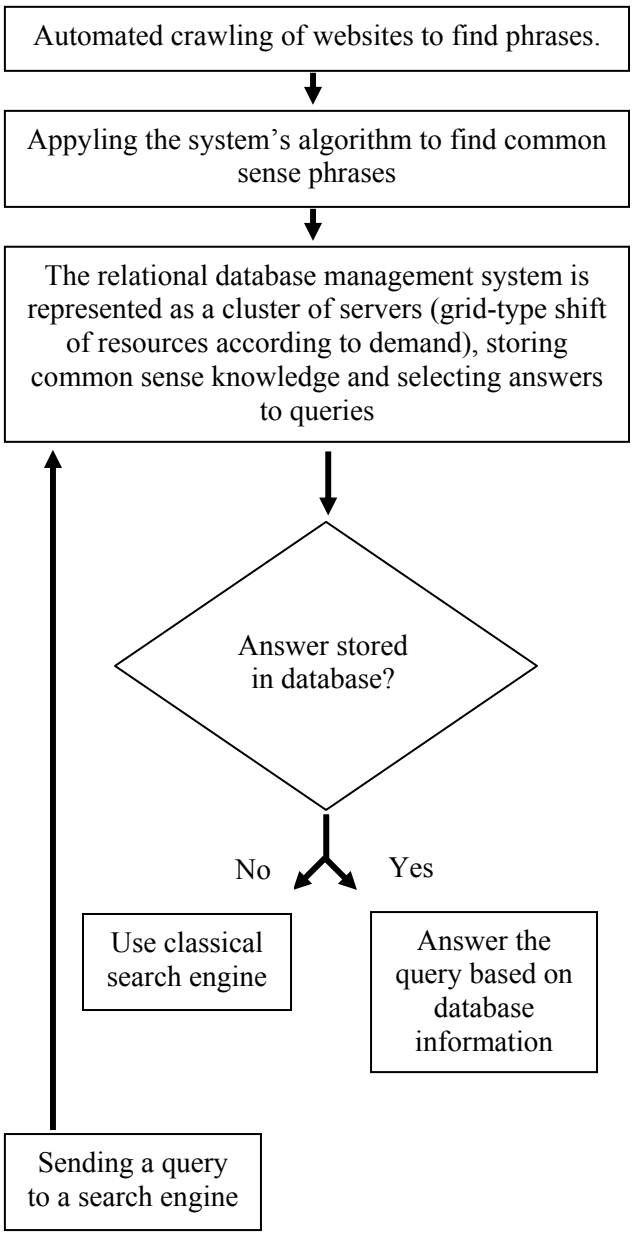

Figure 2: The system as a flow-chart

A flow-chart diagram of the system's individual components is displayed in Figure 2. As can be seen in the flow-chart, the relational database management system fulfills 2 purposes. It stores information and manages queries that come from the search engine by selecting appropriate answers based on the information stored. If there is no relevant answer stored in the database, it will give the answer a classical search engine would provide (e.g. Google). The whole system requires a lot of server resources and is best managed by a grid-computing approach, where resources are shifted according to demand between a cluster of servers.

4 Conclusion

A well-functioning prototype with the aim to improve the quality of search engines has provided some encouraging results. However, it has to be kept in mind that years lie ahead before it is possible to implement a large scale project that replaces present technology. The main reason lies in the expenses of implementing a grid-type cluster of servers that can handle millions of queries and update its knowledge-base at the same time. Moreover, such a project cannot be managed by one researcher alone. The rationale behind this paper is therefore to present the prototype of a system to stimulate interest among other research groups that might wish to collaborate on this goal. The purpose of this project is academic and therefore entirely open-source. It is important to underline that the algorithm used in this paper is just one possible way of tackling this problem. It is probably the case that different methods would solve this problem successfully. One reason in favor of choosing this algorithm was its demonstrated usefulness in terms of simulating human learning and memory [11]-[13]. Likewise, the representation of common-sense knowledge is largely based on human learning and memory. A related approach has been the successful development of a brain-like knowledge navigator for the sciences [15]-[16]. This approach focused on neuroscientific evidence about information processing in the human brain. It diverges from the approach presented here because its goal is context-specific (i.e. focusing on scientific information). Its principles of knowledge representation are general, however, and could thus be applied to search engine technology in general. The same could be true for a number of different approaches that all have the aim to improve search engine technology. The joint effort of combining ideas from various approaches will be an important challenge for the years to come.

References:

[1] N. Zhong, J. Liu, Y. Yao, In Search of the Wisdom Web, *IEEE Computer*, Vol. 35, No. 11, 2002, pp. 27-31.

[2] D. Fensel, Ontology Based Knowledge Management, *IEEE Computer*, Vol. 35, No. 11, 2002, pp. 56-59.

[3] J. Han, K. Chen-Chuan Chang, Data Mining for Web Intelligence, *IEEE Computer*, Vol. 35, No. 11, 2002, pp. 64-70.

[4] N. Cercone, L. Hou, V. Keselj, A. An, K. Naruedomkul, X. Hu, From Computational Intelligence to Web Intelligence, *IEEE Computer*, Vol. 35, No. 11, 2002, pp. 72-76.

[5] L. Zadeh, Keynote lecture given at the FuzzIEEE Conference during the IEEE World Congress on Computational Intelligence, 12 to 16 May 2002, Waikiki, Hawaii.

[6] S. Sekine, R. Grishman, Hindi-English Cross-lingual Question-answering System, *ACM Transactions on Asian Language Information Processing*, Vol. 2, No. 3, 2003, pp. 181-192.

[7] E. Brill, S. Dumais, M. Banko. An Analysis of the AskMSR Question-Answering System. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, University of Pennsylvania, Philadelphia, USA.

[8] V. Chaudhri, R. Fikes, *Question-Answering Systems. Papers from the AAAI Fall Symposium,* AAAI-Press, 1999.

[9] S. Harabagiu, V. Chaudhri, *Mining Answers from Text and Knowledge Bases*, *Papers from the AAAI Spring Symposium*, AAAI-Press, 2002.

[10] http://www.wikipedia.org

[11] R. Spiegel, Human and Machine Learning of Spatio-Temporal Sequences: An Experimental and Computational Investigation. PhD-Thesis, University of Cambridge (UK), 2002.

[12] R. Spiegel, M.E. Le Pelley, M. Suret, I.P.L. McLaren, Combining Fuzzy Rules and a Neural Network in an Adaptive System, *Proceedings of the IEEE International Conference on Fuzzy Systems (FuzzIEEE) in association with the IEEE World Congress on Computational Intelligence*, 2002, pp. 340-345.

[13] R. Spiegel, I.P.L. McLaren, Abstract and associatively-based representations in human sequence learning. *Phil. Trans. R. Soc. Lond. B*, 2003, pp. 1277-1283.

[14] http://www.answerbus.com/

[15] http://www.interactive-systems.de/

[16] M.C. Hirsch, Umsetzung von Text in innere Bilder. Neue Wege in der Informatik. Presentation given at the Kolloquium der Allgemeinen Bildwissenschaften, Burda Akademie, 16 June 2005, Ludwig-Maximilians-University, Munich, Germany.