Ernst-Georg Haffner, Uwe Roth, Christoph Meinel:
A Hyperlink-Proposal Mechanism to Exemplify Cognitive Algorithms for Web-Applications
in Proceedings of the 6th Joint Conference on Information Sciences (JCIS 2002), Durham, North Carolina, USA, pp. 517-520, 3, 2002.

A Hyperlink-Proposal Mechanism to Exemplify Cognitive Algorithms for Web-Applications *

Ernst-Georg Haffner, Uwe Roth, and Christoph Meinel
Institute of Telematics, Bahnhofsstr. 30-32, D-54292 Trier, Germany
{haffner,roth,meinel}@ti.fhg,de

Abstract After a brief description of an algorithm for proposing hyperlinks, this article focuses on a general class of "Cognitive Algorithms" to improve Web-applications. The idea of the hyperlink-proposal mechanism is based on the concept of Case-Based Reasoning, even though there are also several important differences. A closer look at the methodological implementation of this process can also help to find new solutions for other Web-specific problems.

Keywords: Cognitive Algorithms, Hyperlink Proposal, Case-Based Reasoning

1 Introduction

In this paper, we will focus on the research of proposing links for hypertexts in general and a concrete technique based on statistical information retrieval. We will regard the underlying method as an instance of an abstract class of algorithms to improve Web-applications. The idea is to solve the task of finding appropriate links rather fast and straightforward with "learning" and "classifying" phases.

This concept is based on Case-Based Reasoning (CBR) theory, which will be highlighted briefly.

The contribution aims at defining the abstract class of those techniques as *Cognitive Algorithms* for Web-applications. Furthermore, several working areas are enumerated where the same concept may result in important improvements.

2 Hyperlink-Proposal Research

Hyperlinks are most important for HTML documents and they are one reason for the enormous success of the World Wide Web. They create relationships between two documents on base of presenting related work, contradictory positions, further information or simply by the continuation of the next page or by giving pure navigational information [Ric98]. In general, it is rather difficult to find appropriate targets for hyperlinks. The progress made so far on the area of retrieving hyperlinks is summarized in [KKD99].

On the one hand, the best results for proposing hyperlinks are obtained by use of semantic knowledge, but this processing requires enormous user interaction [PMH97]. On the other, approaches on a pure statistical base (see [GZ93]) without any semantic knowledge result in rather low quality proposals [Glu89].

Nevertheless, we focus on retrieving hyperlinks on base of statistical information retrieval with approved methods (e.g. [Teb98]). One interesting concept is, for instance, Chang's *HieNet*. A description of this approach can be found in [Cha93].

In the following, we will describe briefly the generation of hyperlink-proposals on base of some CBR-similar techniques and we will focus on the general processing and the main algorithmic phases. Good examples for classical solutions without using CBR can be found - for instance - in [MS93].

3 Constructing a Hyperlink-Proposal Algorithm

We present a hyperlink-proposal system where all links that are part of any hypertext can be proposed no matter whether the target document is part of the system or not. Storing the possible hyperlinks is not enough, though. Of major importance is the fact that the system must store the relationship between the text and the associated links. This step is called a *learning process*. Therefore, the database can also be called a *knowledge-base*. The quality of this knowledge-base depends on the quality of the learning process: How can the information of the texts be combined with appropriate hyperlinks?

One possibility would be the human teacher. A person or a group of persons could derive the important information of the document manually. This process is very cost- and time-consuming and is not feasible in practice, especially in the area of Web-applications. Most of today's learning algorithms conquer the problem of high quality knowledge retrieval by extracting some information automatically

* Proceedings of the 6th Joint Conference on Information Sciences (JCIS 2002), March 2002, NC, USA, pp. 517-520.

on the basis of advanced heuristics. The learning process itself evaluates the relevance of the extracted information.

At first, the learning algorithm treats a text with hyperlinks as if it does not contain any link and derives the relevant information. Next, it proposes one or more hyperlinks and compares the result to the hyperlinks that the document in fact contains. By using this method, the learning process can be carried out without any human teacher. A disadvantage arises from the strong relationship between the quality of the learning process and the quality of the initial documents. Furthermore, only those hyperlinks can be proposed that are already known to the system and at the very beginning of the learning process there are no links to be proposed at all.

We modeled the hyperlink generation problem as a kind of Case-Based Reasoning system (CBR). Research in the area of Case-Based Reasoning (CBR) began in the early years of the last decade [Aha91]. CBR-systems are well known means of representing knowledge in form of *cases*. Each case can be regarded as a *problem* together with its *solution*. A problem consists of its description in form of *attributes* and one or more solutions which refer to it.

We used the CBR-concept to retrieve high quality links as proposals for the web author. The written texts of the web authors are regarded as the *problems* and the hyperlinks within are the *solutions*. A complete hypertext can thus be viewed as a *case*. In the learning process, (statistical) text attributes are stored together with their attached hyperlinks into the knowledge-base, in the CBR-environment also called the *case-base*. In the classifying step, raw texts are presented to the system which proposes hyperlinks for the text as solutions. A possible *a posteriori* classification to evaluate the quality of the proposals is illustrated in Figure 1.

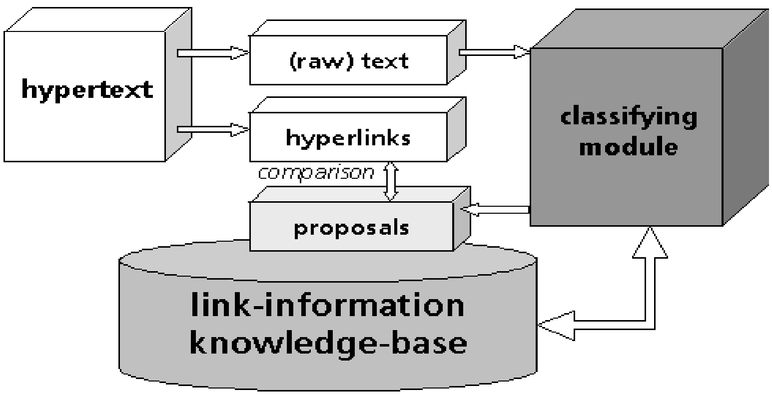

Figure 1. Learning and classifying applied to the hyperlink context

Furthermore, every new hypertext generates an additional solution - namely the link that refers to itself. We designed the CBR-model for the hyperlink environment with further differences to the classical approach. Further details on the implementation of the system can be found in [HHR00].

4 Evaluation Highlights

We tested the described system on base of a large amount of data “a posteriori” on existing web pages. The advantage of this processing is the “real-life” usability of the approach.

To evaluate the quality of the HPM, we extracted the links within several HTML-files, classified the raw texts using our model, and finally compared the classification results to the existing hyperlinks according to the proposed learning and classifying steps.

As measurement for the evaluation step, we introduced new terms on base of the probabilities of the link proposals, namely the *Quantified Cumulating Recall* (QCR) and the *Quantified Cumulating Precision* (QCP). The QCR is derived from the term *recall*, that denotes the share of appropriate links among all “good links” of the hypertext. The gradient of that curve signals the recall-share of the proposals as one measurement of the quality of the HPM results. For more information on the QCP and further evaluation criterions see [HHR00].

The graphical QCR-results of the approach for the "Association for Computing Machinery" (ACM), the "World Wide Web Consortium" (W3C), and the "Association for the Advancement of Computing in Education" (AACE) are shown in figure 2.

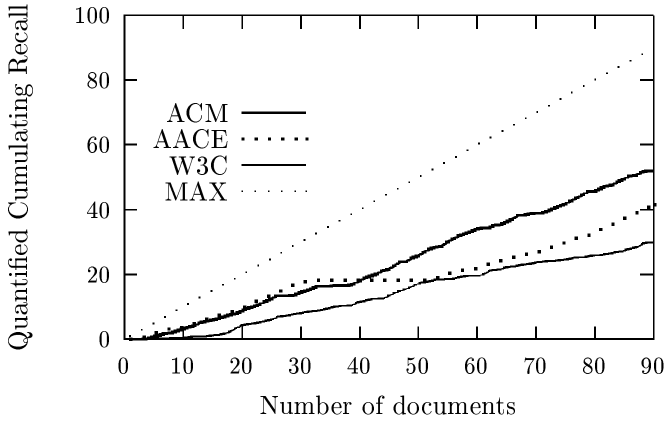

Figure 2. Quantified Cumulating Recall (ACM, AACE, W3C)

5 Cognitive Algorithms

An abstraction and further generalization of the process to propose hyperlinks leads to the concept of *Cognitive Algorithms* to improve Web-applications. Their main characteristic is the distinction between *learning* and *classifying* phases. In Computer-Science, there are several different basic concepts and methods for the design of algorithms. A well-known technique is for instance *Divide-and-Conquer*. The general idea of this methodology is to divide a complex problem into several at least slightly simpler sub-problems.

To describe and classify the well-known techniques approved in the area of AI we will use terms of *cognitive science*. A central aspect hereof is *knowledge*. Knowledge retrieval and storage in knowledge bases is regarded as a kind of *learning*.

Certainly, not every algorithm that stores a "1" in a database "learns". Learning can happen by acquiring knowledge, by recognizing legalities or rules, or by analogous deduction where the similarity plays an important part.

In fact, the term of "learning" should only be applied to systems that are able to transform complex, unstructured, incomplete, inconsistent or incorrect input into simple, well structured, complete, consistent or error free output based on a dynamically changing knowledge base. This transformation can be called *classification*. Expert systems, for instance, fulfill most of the described requirements.

Expert systems can be *analytic* or *synthetic*. A good example of the former is a *diagnose system* that works on the base of existing data while a *planing system* that is able to deduct new data corresponds to the latter.

Mostly, expert systems are very large, complex programs that have to be attended and maintained during a long time span in order to become absolutely reliable. They can store the knowledge of human experts either explicitly or implicitly. Detailed information on the whole topic of AI-principles in general and especially on expert systems can be found in [Ric89].

The idea to propose hyperlinks with expert systems might be appropriate. A synthetic modeling would be able to generate new links on base of the textual information while an analytical modeling would propose links that were learnt before. Even though our described modeling cannot be regarded as an expert system due to its simplicity and lack of some needed properties, it is an analytic tool that performs learning and classification tasks.

We call algorithms that can be described adequately in terms of learning and classifying as *Cognitive Algorithms*. There are several further Web-applications, which might be appropriate for this kind of modeling. Certainly, the main ideas and concepts that were used to implement straightforward solutions for hyperlink-proposals can also be applied to the following examples:

– *Intelligent mail filters*

Here, the learning process consists of recognizing user action: Which mails belong to which folders? The classification step would result in automatically structuring incoming mails into the corresponding folders.
- *Web searching algorithms*
  An area where *Intelligent Agents* or *Mobile Agents* are well suited is web searching. They, too, can belong to the class of Cognitive Algorithms. Learning means understanding the meaning of the user requests while the classification results in consistent search outcomes.[2]
- *Intelligent data routing*
  The Institute of Telematics is currently working on an advanced Smart Data Server project with intelligent data transfer routing for the Web. The idea of considering the actual server loads to re-route certain client requests can be modeled with Cognitive Algorithms. The learning consists in measuring current response-times and deducing the reasons. The classification would result in an efficient request routing.

The strategy of modeling Cognitive Algorithms leads to potential improvements of manifold web-applications due to synergy-aspects. Furthermore, this synergy does not only consist of sharing algorithms for data manipulation and complex objects with general usable methods, but also of higher modeling levels. Thus, ideas to improve one Cognitive Algorithm can - potentially - be an advantage for most of the others too.

6 Summary and Conclusion

In this paper, we sketched the idea to transfer concrete concepts for proposing hyperlinks to other areas of Web-applications.

The "learning" and "classifying" phases were derived from Case-Based Reasoning research. We generalized those methods and called the abstract class "Cognitive Algorithms". For the generation of hyperlink proposals, the described methodology was rather promising.

We think that several further concept-improvements in the area of Cognitive Algorithms will lead to more efficient and faster Web-applications in the future and, thus, to an improvement for the Web community.

References

[Aha91] David W. Aha. *Case-Based Learning Algorithms.* Proceedings of the DARPA Workshop on Case Based Reasoning, Morgan Kaufmann, 1991. 147-157

[Cha93] Daniel T. Chang. *HieNet: A User-Centered Approach for Automatic Link Generation.* Proceedings of the Fifth ACM Conference on Hypertext, Hypertext '93, ACM, 1993. 145-158

[EK98] Elizabeth A. Kendall, P. V. Murali Krishna, Chirag Pathak, C. B. Suresh. *Patterns of Intelligent and Mobile Agents.* Proceedings of the Second International Conference on Autonomous agents, ACM, 1998. 92-99

[Glu89] Robert J. Glushko. *Design Issues for Multi-Document Hypertexts.* Proceedings of the 2nd ACM Conference on Hypertext, Hypertext'89, ACM, Pittsburgh, PA, 1989. 51-60

[GZ93] J. Gordesch, A. Zapf. *Computer-Aided Foramtion of Concepts.* L. Hrebicek, G. Altmann (eds.). Quantitative text analysis, Quantitative linguistics, Vol. 52, WVT Trier, 1993

[HHR00] Ernst-Georg Haffner, Andreas Heuer, Uwe Roth, Thomas Engel, Christoph Meinel. *Advanced Studies on Link- Proposals and Knowledge-Retrieval of Hypertexts with CBR.* Proceedings of the International EC-Web Conference, ECWeb2000, Greenwich, United Kingdom, Springer-Verlag LNCS 1875, 2000. 378-396

[KKD99] Hermann Kaindl, Stefan Kramer, Papa Samba Niang Diallo. *Semiautomatic Generation of Glossary Links: A Practical Solution.* Proceedings of the Tenth ACM Conference on Hypertext, Hypertext '99, ACM, 1999. 3-12

[MS93] C. Marshall, F. Shipman. *Searching for the Missing Link: Discovering Implicit Structure in Spatial Hypertext.* Proceedings of the Fifth ACM Conference on Hypertext, Hypertext '93, ACM, 1993. 217-230

[PMH97] C. Petrou, D. Martakos, S. Hadjiefthymiades. *Adding Semantics to Hypermedia Towards Link's Enhancement and Dynamic Linking.* Hypertext - Information Retrieval - Multimedia '97, HIM 1997, Universitaetsverlag Konstanz, 1997

[Ric89] Michael M. Richter. *Prinzipien der Künstlichen Intelligenz.* B. G. Teubner Stuttgart, 1989

[Ric98] F. J. Ricardo. *Stalking the Paratext: Speculations on Hypertext Links as Second Order Text.* Proceedings of the Ninth ACM Conference on Hypertext, Hypertext '98, ACM, 1998. 142-151

[SW86] Craig Stanfill, David Waltz. *Toward Memory-Based Reasoning.* Communications of the ACM, 29 (12), 1986. 1213-1229

[Teb98] John Tebbutt. *Finding links.* Proceedings of the Ninth ACM Conference on Hypertext, Hypertext '98, ACM, 1998. 299-300

[2] Further information on mobile or intelligent agents can be found in [EK98], for instance.