Pergamon

Information Processing & Management, Vol. 33, No. 2, pp. 129–131, 1997
© 1997 Published by Elsevier Science Ltd
All rights reserved. Printed in Great Britain
0306-4573/97 $17+0.00

PII: S0306-4573(96)00057-X

INTRODUCTION TO THE SPECIAL ISSUE ON METHODS AND TOOLS FOR THE AUTOMATIC CONSTRUCTION OF HYPERTEXT

MARISTELLA AGOSTI[1] and JAMES ALLAN[2]

[1] Department of Electronics and Informatics, University of Padova, Italy and [2] Department of Computer Science, University of Massachusetts—Amherst, MA 01003, U.S.A.

The tremendous popularity of the *World Wide Web* has created a corresponding demand for on-line data organized as a hypertext or hypermedia document collection. To date, most of that organization has been done by hand, a daunting—if not impossible—task for very large or very volatile collections such as archival collections or newswire services. To help organize such collections (as well as those that are smaller or less dynamic) techniques are required for automatically constructing hypertext or hypermedia, or minimally for providing user-assistance in that process.

When considering a method for automatic hypertext link construction, many aspects need to be addressed, because the construction of a hypertext that is to facilitate the user in retrieving useful information, needs to make available both the functionalities and capabilities of a *hypertext system* together with those of an *information retrieval system.* That is, it is necessary to create a structure that gives the user the combination of *browsing*, and *searching* facilities, not simply one without the other. A *hypertext information retrieval* (HIR) system can be made available to the user as a tool that combines both the sophisticated and advanced searching capabilities that have been developed in modern IR systems together with the navigation and browsing facilities of hypertext systems. (Agosti & Smeaton, 1996).

Since a crucial part of the automatic construction of a hypertext is the creation of links connecting documents or document fragments that are semantically related, some Information Retrieval researchers have begun using IR techniques for that automatic construction, because IR has always dealt with the construction of relationships between objects that are relevant to similar topics—or to each other.

This issue addresses different approaches for the automatic transformation of collections of "flat" textual documents to produce a structured hypertext. Tools and methods capable of producing an informative hypertext collection of documents that can be searched and browsed by content are addressed, together with techniques for automatically augmenting the links in an existing hypertext and for evaluating the hypertext automatically built.

THE PAPERS IN THIS ISSUE

Although the problem of automatic hypertext construction has become particularly significant only within the past few years, researchers have been addressing aspects of the problem for quite some time. The paper by Agosti, Crestani, and Melucci, *On the use of Information Retrieval Techniques for the Automatic Construction of Hypertext*, presents a survey of what has been achieved, and what have been the main and most successful lines of research.

The paper by Allan, *Building Hypertext Using Information Retrieval*, proposes a methodology for automatic hypertext construction that makes use of the vector space model of information retrieval for automatically linking related documents—and most important and novel, a process for automatically assigning a type to some documents relationships. The types that are found automatically are: revision, summary, expansion, equivalence, comparison, contrast, tangential, and aggregate links. The paper raises the point that statistical information retrieval techniques are imperfect and unable to handle fine distinctions in meaning, so these techniques work well in general settings, but some human intervention can be necessary in really specific domains where greater accuracy can be required. An initial evaluation of the work is also included.

Thistlewaite, in *Automatic Construction and Management of Large Open Webs*, addresses the important issue of dynamic identification and creation of hypertext links in an open hypertext environment, where documents are being added constantly and where fixed links are inadequate. The solution proposed is based on pattern recognition in text strings for link construction. Most of what is discussed has been tested in a solid implementation which creates a hypertext from the complete electronic document holdings of the Australian Parliament.

In *Browsing Document Collections: Automatically Organizing Digital Libraries and Hypermedia using the Gray Code*, Losee presents an adaptive model that can be used to automatically organize documents in hypermedia systems and digital libraries. The model is based on the Gray code that is used to classify and order the documents. In particular the ordering principle is used to place similar documents adjacent to each other, providing a clustering of related documents of the same sort as that provided by hypermedia links.

Salton, Singhal, Mitra and Buckley in *Automatic Text Structuring and Summarization* face the problem of identifying internally consistent text pieces from available texts, where the main target is the use of more meaningful text units than the usual physical text elements. The specific goal addressed is the development of a complete summarization scheme to automate the process. The work proposed has been evaluated by comparing automatically generated extracts with those produced by a group of experts. We are particularly pleased to include this paper because it covers much of the work that Professor Salton was studying before his death in the summer of 1995. Some aspects of the work deserve further study and we hope that publishing the paper will inspire such investigation.

The paper by Lewis and Knowles, *Threading Electronic Mail: A Preliminary Study*, addresses the problem of linking e-mail messages or electronic news articles with others in the same "thread" of discussion. The paper starts from a real application problem and shows possible means for automatically linking different textual documents (e-mail messages) that are pertinent to each other.

The work by French, Knight and Powell, *Applying Hypertext Structures and Software Documentation*, discusses their SLEUTH system, which includes automatic linking to help with the problem of managing software documentation.

Tudhope and Taylor in *Navigation via Semantic Similarity: automatic linking based on semantic closeness*, present another application of automatic hypermedia construction based on a collection of historical photographs together with textual and oral histories. The paper provides interesting reading about a significant real application and is one of the few papers which addresses issues related to similarity of images or locations.

Hammwöhner and Rittberger, in *Building application dependent hypertexts*, face again the application facet, but the system presented in this paper is related to the use of the Konstanz Hypertext System (KHS) for the automatic production and construction of large, multi-disciplinary and multi-functional hypertexts.

The final paper of the issue, *Methods for Evaluating the Quality of Hypertext Links*, is by Blustein, Webber and Tague-Sutcliffe. It addresses the fundamental facet of evaluation of automatically constructed hypertexts and hypermedia. The problem of evaluation is one of the more difficult questions raised in the automatic construction of hypermedia. The IR community has a strong tradition of testing the efficacy of retrieval techniques using standard test collections and effectiveness measures, but standard collections for the evaluation of these new types of document links do not exist. The technique proposed by the authors is to have a "correct" hypertext and to compare resulting hypertexts to that one by examining coverage of generated links.

BACKGROUND

This special issue was originated by the workshop, *Information Retrieval and Automatic Construction of Hypermedia*, held in July 1995 in Seattle in conjunction with the ACM SIGIR

'95 Conference (Agosti & Allan, 1995).

The researchers from academia and industry that met in the workshop were focusing their discussions on how Information Retrieval might help solve the problem of the automatic construction of hypertext and the presented papers were discussing work in various related aspects.

The workshop highlighted the importance and difficulty of the problem and indicated the value of more in-depth study. To address different aspects related to this problem and help consolidate quality research on it, this special issue was originated.

REFERENCES

Agosti, M., & Allan, J. (1995). *IR and Automatic Construction of Hypermedia.* In E. A. Fox, P. Ingwersen, and R. Fidel (Eds), Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle (USA), 379.

Agosti, M., & Smeaton, A. (Eds) (1996). *Information Retrieval and Hypertext.* Kluwer Academic Publishers, Boston.