ELSEVIER

Decision Support Systems 35 (2003) 1–5

Decision Support Systems

www.elsevier.com/locate/dsw

Editorial

Special issue: "Web retrieval and mining"

1. Introduction

Search engines and data mining are two research areas that have experienced significant progress over the past few years. Overwhelming acceptance of the Internet as a primary medium for content delivery and business transactions has created unique opportunities and challenges for researchers. The richness of the web's multimedia content, the reach and timeliness of web-based publication, the proliferation of e-commerce activities and the potential for wireless web delivery have generated many interesting research problems. Technical, system, organizational and social research approaches are all needed to address these research problems. Many interesting web-retrieval and mining research topics have emerged recently. These include, but are not limited to, the following:

- Text and data mining on the web
- Web visualization
- Web intelligence and agents
- Web-based decision support and knowledge management
- Wireless web retrieval and visualization
- Web-based usability methodology
- Web-based analysis for eCommerce applications

2. Academic roots: information retrieval, artificial intelligence and web computing

Web retrieval and mining owes much of its success to three areas of research: information retrieval (IR), artificial intelligence (AI) and web computing (WC).

2.1. Information retrieval for web retrieval and mining

Salton [8], a pioneer in IR since the 1970s, is generally considered to be the father of IR. His vector space model has become the foundation for representing documents in modern IR systems and web search engines.

IR is a field that has gone through several major generations of development. In the 1970s, computational techniques based on inverted index and vector space were developed and tested in computer systems. In addition, Boolean retrieval methods and simple probabilistic retrieval models based on Bayesian statistics were created. Although more than 30 years old, this set of techniques still forms the basis of modern IR systems and Internet search engines.

In the 1980s, coinciding with the developments of new AI techniques, knowledge-based and expert systems were developed that aim to emulate expert searchers and domain specialists. User modeling and natural language processing (NLP) techniques were developed to assist in representing users and documents. Research prototypes were created that employee production systems representation and heuristics for effective online searching. Most Internet search engines and web portals have incorporated some simple forms of NLP.

Realizing the difficulties in creating domain-specific knowledge bases and heuristics, researchers in the 1990s attempted to adopt new machine learning techniques for information analysis. AI techniques, including neural networks, genetic algorithms and symbolic learning, were tested in IR [2]. Many of these techniques have been adopted successfully in data mining [4,7] and, more recently, in text mining [3].

0167-9236/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-9236(02)00104-5

Since the mid 1990s, the popularity of search engines and advances in web spidering, indexing and link analysis [1] transform IR systems into newer and more powerful search tools for locating content on the Internet. The multimedia content and the ubiquitous presence of the web allow both commercial users and the general public to see the potential for utilizing unstructured information assets for their everyday activities and business decisions.

2.2. Artificial intelligence for web retrieval and mining

Herbert Simon is generally considered one of the founding fathers in artificial intelligence (AI). The field has long been aiming to model and represent human intelligence in computational models and systems [6].

Simon et al. have pioneered the early works in AI, most notably the General Problem Solvers (GPS) that emulated general human problem solving. In the 1970s, computer programs were developed to emulate rudimentary but human-like activities such as crypt-arithmetic, chess, game, puzzle, etc.

In the 1980s, there was an explosion of AI research activities, most notably in expert systems. Many research prototypes were created to emulate expert knowledge and problem solving in domains such as medical and car diagnosis, oil drilling, computer configuration, etc. However, the failure of many of such systems in commercial arenas tarnished the name of AI for some years. Many venture capitalists have backed away from any ventures having association with AI.

The many failures of commercial expert systems have nevertheless made both researchers and practitioners become realistic about the strengths and weaknesses of such systems. Expert systems are not silver bullets, instead, they are suited for well-defined domains with willing experts. Domain-specific heuristics and language parsing rules also appear to be useful for web retrieval and mining applications.

In the 1990s, AI-based symbolic learning, neural networks and genetic programming generated many significant and useful techniques for both scientific and business applications. The field of data mining is the result of significant research developed in this era. Many companies have since applied such techniques in successful fraud detection, financial prediction and customer behavioral analysis applications [4]. Web mining inherits many similar and useful data analysis techniques from data mining research.

Both IR and AI research have provided a foundation for knowledge representation. For example, indexing, subject headings, dictionaries, thesauri, taxonomies and classification schemes are some of the IR knowledge representations that are widely used in various knowledge management practices, while AI researchers have developed knowledge representation schemes such as semantic nets, production systems, logic, frames and scripts.

2.3. Web computing for web retrieval and mining

Web computing is a new discipline that had emerged from the significant research opportunities and developments in the areas of network infrastructure, protocol, software engineering and algorithmic research related to both the Internet and the World Wide Web (WWW).

Web computing is based on significant network protocol research such as TCP/IP, Ethernet and http. Vint Cerf and Tim Berners-Lee are generally considered the fathers of the Internet and WWW, respectively. Research in pre-WWW days laid the foundation for modern networking and communication architecture and standards.

Since the web infrastructure became robust, web retrieval (of information instead of interconnection of computers) has become a major part of everyday life. Popular search engines, from Excite and Google to Alta Vista and Yahoo, have created many ubiquitous “digital libraries” of web content through their indexing, ranking, spidering and searching technologies [1]. However, these search engines are known to have suffered with problems of information overload, low precision, information quality and information coverage [5]. The saddest consequence, however, is that most search engines have deviated from their technological root and become media companies (that stress content and advertising).

Web mining, we believe, is still in its infancy. While the disciplines of data mining and web retrieval (search engines) are becoming relatively mature, web mining has not yet converged into a discipline of known techniques and practices. However, many interesting research topics and approaches to addressing these topics have emerged: from search patterns to web business transactions, and from analytical techniques to visualization of web mining results. The next 5–10 years promise to be an exciting time for researchers in this area.

3. In this issue

This special issue consists of nine papers that report research in web retrieval and mining.

The first paper, "EDGAR-Analyzer: Automating the Analysis of Corporate Data contained in the SEC's EDGAR Database," by Gerdes, presents a tool (EDGAR-Analyzer) that automates the analysis of SEC filings, with particular emphasis on the unstructured text sections of these documents. The results of a large-scale case study of corporate Y2K disclosures in 18,595 10 K filings are presented. The second paper, "Enhancing the Power of Web Search Engines by Means of Fuzzy Query," by Choi, proposes a new measure called the Perception Index (PI) that contains attributes associated with a focal keyword restricted by fuzzy terms used in fuzzy queries on the Internet. The new measure allows the expression of preferences and vagueness for Internet queries that are not supported by current search engines. The third paper, "Feature Selection on Hierarchy of Web Documents," by Mladenic and Grobelink, describes the use of feature subset selection on text data. In their experiments, naïve Bayesian classifier was used to help predict probability that a new example is a member of the corresponding category. The fourth paper, "Visualization of Large Category Map for Internet Browsing," by Yang, Chen and Hong, attempts to address the information overload problem of the web by employing visualization techniques. Their results show that both fisheye views and fractal views significantly increase the effectiveness of visualizing category map for web collections. The fifth paper, "Integrating Web-based Data Mining Tools with Business Models for Knowledge Management," by Heinrichs and Lim, investigates the relationship between the independent variables of web-based presentation and data mining tools and business models and the dependent variable of strategic performance capabilities. Their results demonstrate a positive interaction effect between the variables. The sixth paper, "Enriching Web Taxonomies through Subject Categorization of Query Terms from Search Engine Logs," by Chuang and Chien, proposes a query categorization approach to facilitate the process of constructing web ontologies. Their technique categorizes web query terms from the search logs into a predefined concept taxonomy based on popular search interests. The seventh paper, "Automatic Information Extraction from Semistructured Web Pages by Pattern Discovery," by Chang, Hsu and Lui, proposes a pattern discovery approach that can extract structured data from semistructured web documents. Their prototype system (IEPAD) applies several pattern discovery techniques, including PAT trees, multiple string alignments and selected pattern matching algorithms. The eighth paper, "Automatic Discovery of Similarity Relationships through Web Mining," by Roussinov and Zhao, shows how the web can be mined in an automated manner to discover semantic similarity relationships among concepts that surfaced during an electronic brainstorming session. The paper develops a new method called Context Sensitive Similarity Discovery for mining similarity relationships based on the Organizational Concept Space proposed by the authors in their earlier research. The ninth and last paper, "Design and Evaluation of a Multiagent Collaborative Web Mining System," by Chau, Zeng, Chen, Huang and Hendriawan, presents the *Collaborative Spider*, a multiagent system designed to provide postretrieval analysis and enable across-user collaboration in web search and mining. Their experiments show that subjects' search performance was degraded when they had access to a limited number (i.e., 1 or 2) of earlier searches. However, search performance improved significantly when subjects had access to more search sessions.

We hope this collection of research papers will help advance our knowledge and understanding of this fascinating and evolving field of web retrieval and mining.

4. The future

What are the future research areas for web retrieval and mining? What are the applications and technologies that may affect future knowledge workers and decision makers?

4.1. Semantic Web

The current web infrastructure is one of hypertext syntax and structure. The html hyperlinks do not suggest any semantic or meaningful relationships between web pages. We only know that two hyperlinked pages have some sort of relationship. How to represent the semantics on the web and to create a Semantic Web of meaningful interconnected web objects and content is a challenging research topic. Some researchers suggest richer web content notations and inference rules using representations such as XML and RDF, others suggest a system-aided machine learning approach to extracting semantic associations between objects.

4.2. Multilingual Web

The web has increasingly become more international and multicultural. Non-English web content has experienced the strongest growth over the past few years. In addition, the trend in globalization and e-commerce has created voluminous multilingual Intranet content for multinational corporations or companies with international partners. How can we create a multilingual knowledge portal such that users can experience seamless cross-lingual information retrieval (e.g., searching for Chinese government regulations using English queries) and real-time machine translation? The most immediate application of a multilingual web would be in international marketing and intelligence analysis for multinational corporations.

4.3. Multimedia Web

We believe the trend toward multimedia data mining cannot be reversed. Although it may not become a large part of corporate information assets (unlike structured data and unstructured text), it does fill an important gap in corporate knowledge management.

4.4. Wireless Web

Although we believe most web content will continue to be accessed over a high-speed wired network, wireless applications will continue to proliferate in years to come. They will also emerge and proliferate rapidly in selected high-impact application areas, e.g., email, financial stock quotes, curbside law enforcement alerting (via PDA and cell phone), etc. For knowledge workers who are mobile and time-pressed, wireless web retrieval and mining will not be a luxury but a necessity in a not-so-distant future.

References

[1] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, The 7th WWW Conference, 1998, http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm.

[2] H. Chen, Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms, Journal of the American Society for Information Science 46 (3) (April 1995) 194–216.

[3] H. Chen, Knowledge Management Systems: A Text Mining Perspective, The University of Arizona, Tucson, AZ, 2002.

[4] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, Menlo Park, CA, MIT Press, Boston, MA, 1996.

[5] S. Lawrence, C.L. Giles, Accessibility of Information on the Web, Nature (400) (1999) 107–109.

[6] A. Newell, H. Simon, Human Problem Solving, Prentice-Hall, Englewood Cliffs, NJ, 1972.

[7] I.H. Witten, E. Frank, Data Mining, Morgan Kaufmann, San Francisco, CA, 2000.

[8] G. Salton, Automatic Text Processing, Addison-Wesley, Reading, MA, 1989.

Dr. Hsinchun Chen is McClelland Endowed Professor of MIS at The University of Arizona (fourth-ranked in the field of MIS according to the US World and News Report). He received his PhD degree in Information Systems from New York University in 1989. He is author of more than 100 articles covering medical informatics, semantic retrieval, search algorithms, knowledge discovery and collaborative computing in leading information technology publications. He serves on the editorial board of *Journal of the American Society for Information Science and Technology* and *Decision Support Systems*. Dr. Chen founded The University of Arizona Artificial Intelligence Lab in 1990. As a major research group within the university, the Artificial Intelligence Lab employs over 40 staff, research scientists, research assistants and programmers and has made a significant contribution to the research and educational experience of students at all levels in the MIS Department of The University of Arizona. Since 1990, Dr. Chen has received more than $12 M in research funding from various government agencies and major corporations including NSF, DARPA, NIJ, NIH, NLM, NCI, HP, SAP, 3COM and AT&T. Dr.

Chen is the Founding Director of The University of Arizona Mark and Susan Hoffman E-Commerce Lab (October 2000), which features state-of-the-art hardware and software in a cutting-edge e-commerce and enterprise computing research and education environment. Dr. Chen is also founder of a knowledge management technology and service company—Knowledge Computing Corporation, a University of Arizona spin-off company. The company, which is Tucson-based, had received major venture capital funding and is growing rapidly in the law enforcement and market portal sectors. Dr. Chen's work also has been recognized by major US corporations for his contribution to IT education and research. In 1995 and 1996, he received the AT&T Foundation Award in Science and Engineering. In 1998, he received the SAP Award in Research/Applications and became the Karl Eller Center Honored Entrepreneurial Fellow. In 1999, Dr. Chen received the McClelland Endowed Professorship and the Andersen Consulting Professor of the Year Award. In 2000, he received the Kalt Prize for Doctoral Placement. Dr. Chen has been heavily involved in fostering digital library and knowledge management research and education in the US and internationally. He was a PI of the NSF-funded Digital Library Initiative-1 project and he has continued to receive major NSF awards from the ongoing Digital Library Initiative-2, ITR and Digital Government programs. He also helped organize and promote the Asian digital library research community and has served as either the conference general chair or international program committee chair for five International Conferences of Asian Digital Libraries (ICADL; in Hong Kong in 1998, in Taipei, Taiwan in 1999, in Seoul, Korea in 2000, in Bangalore, India in December 2001 and in Singapore in 2002). Dr. Chen has frequently served as a panel member and/or workshop organizer for major NSF research programs. Dr. Chen is also a recognized advisor for international IT research programs in Hong Kong, China, Taiwan, Korea and Ireland.

Hsinchun Chen
Artificial Intelligence Laboratory and
Hoffman E-Commerce Laboratory,
Management Information Systems Department,
The University of Arizona,
Tucson, AZ 85721, USA
Email address: hchen@eller.arizona.edu