Association for Information Systems

AIS Electronic Library (AISeL)

ICEB 2003 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-9-2003

An Intelligent Web Index Engine for Financial Articles

Drew Hwang

Che-peng Lin

Follow this and additional works at: https://aisel.aisnet.org/iceb2003

This material is brought to you by the International Conference on Electronic Business (ICEB) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICEB 2003 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

An Intelligent Web Index Engine for Financial Articles

Drew Hwang
Computer Information Systems Department
California State Polytechnic University
Pomona, California, USA
dhwang@csupomona.edu

Che-peng Lin
Department of Business Administration
National Changhua University of Education
Changhua, Taiwan 500
cplin@cc.ncue.edu.tw

Abstract

The Web has become a major repository for financial articles, and the indexing of these articles in an efficient and intelligent manner has becoming a vital task. This paper describes an intelligent index engine developed to directly log-on to financial Web sites, download financial articles, parse the texts, and index the articles via the use of a semantic network in the domain of corporate economics. The system can be used for retrieving relevant financial information for corporate performance analysis.

1. Introduction

In corporate performance analysis, a successful financial analyst not only observes numbers shown on financial statements, cash flow schedules, or economic index summaries, but also examines related textual information contained in corporate reports, trade journals, government bulletins, and business newspapers that are mainly descriptive and qualitative in nature. Today, the Web has become a major repository of such qualitative information, and the indexing of financial articles from the Web in an efficient and intelligent manner has becoming a vital task.

This paper describes the design of an intelligent indexing engine for Web financial articles in the domain of corporate performance analysis. The paper first discusses characteristics of the domain, a model that can be used to effectively capture and represent the knowledge in the domain, and semantic networks developed to structure and store the domain knowledge. The paper then presents the design of the Web indexing engine, followed by an illustration of how the system is used. At the end, the paper concludes with recommendations for future research and development.

2. The domain

Similar to evaluating the health of a patient, corporate performance analysis involves a process of diagnosing the state of the health of the underlying company, and the effectiveness of such a diagnosis process relies on support of relevant and timely information. In this domain, the use of quantitative method such as financial ratio analysis is common. However, financial analysts must open their analyses not only for quantitative constructs in financial statements, but also for many qualitative measures, such as marketing strategies, product development, and so on, contained in corporate reports, trade journals, government bulletins, business newspapers, and so on.

The domain of corporate economics includes tremendous amounts of knowledge with various types of contents and appearances, which makes it essential to find an appropriate representation scheme to structure the knowledge. Valid knowledge representation schemes should be built on common theoretical foundation that can be easily understood by the people in the industry.

Traditional financial reporting system is grounded on the value-driven accounting approach. This approach, however, leads to a very rigid definition for an organization's economics model. The counter-approach to the valued-driven approach is the event-driven accounting theories which advocate that an enterprise's approach to corporate information management should focus on managing relevant business events as opposed to managing values shown on the financial reports. The event-driven approach was later applied to develop an entity relationship view of corporate economics called the REA (Resources, Events, and Agents) model [3,4,6]. The model is composed of three classes of economic objects such as resources, events, and agents, and relationships among those objects.

Intelligent information retrieval systems frequently use semantic networks to provide knowledge support for textual information processing for domains with deep knowledge [1]. The concept of the REA modeling technique can serve as the foundation for the development of a semantic network to model knowledge of corporate economics for two reasons. Firstly, the REA model has been proved to be a reliable and consistent knowledge representation scheme for corporate economics. Secondly, because its representation of entity-relationships are exclusively in one-to-one form, the concept of REA model can be used and expanded to develop a semantic network which also uses binary form to represent knowledge.

3. The semantic network

A semantic network is a directed graph in which knowledge objects are the nodes connected by their declared relationships as the links. In the semantic network of corporate economics, the three REA classes of economic objects are the knowledge objects. Defined and identified knowledge objects are related by three sets of binary semantic relationship: generalization, association, and causality. Paired knowledge objects with the semantic relationship are expressed in terms of three semantic predicates: IS_A (generalization), IS_PART_OF (association), and IS_CAUSED_BY (causality). For instance, a "salesman" IS_A "sales representative"; “equipment” IS_PART_OF “asset”; "sales” IS_CAUSE _BY “marketing".

Most semantic networks use the "spreading activation" search algorithm, which navigates the knowledge base by first activating the key objects, followed by objects that are directly related to the objects, followed by objects that are directly related to the objects previously activated, and so on [2]. Psychologists have long recognized that human memory is a system of associated concepts or "chucks" of symbolic information" [5]. The structure of semantic network is analogous to that of the human memory, and the mechanism of spreading activation mimics how human retrieve memory.

Cohen and Kjeldsen [2] also propose three rules that can be used to guide the search. First, the search can cease at a predetermined “distance” (e.g., 3 links or 4 nodes) from the original activation. Secondly, a “fan-out” algorithm can be used to stop the search at objects that have very high connectivity, or fan-out. The third type of constraint uses the idea of likelihood of the relationship between objects. In Cohen and Kjeldsen's [2] GRANT, an intelligent information retrieval system, a minimally constrained spreading activation of a given distance (4 links) is used with satisfactory rates of precision and recall.

4. The system

The intelligent index engine is written in Visual Basic 6 using MS SQL server for data management. The system contains four modules: Request, Index, Retrieve, and Administration.

4.1. The Request module

The Request module uses a COM object to log-on to a financial Web site and downloads a specified HTML page via the HTTP protocol. The module can either download one page at a time or conduct a batch download based on a list of URLs stored in the database.

4.2. The Index module

The Index module parses the downloaded HTML page, analyzes the text, indexes the articles via the use of the semantic network of corporate economics, and saves the article. The module reads every word of the article, matches the word with the economic objects stored in the semantic network, and build indexes of the article accordingly.

4.3. The Retrieve module

The Retrieve module enables users to retrieve relevant articles based on derived keywords and specified distance for spreading effect. There are two processing elements involved in the derivation of search keys: the semantic network and a search key generator. The semantic network discussed previously is the knowledge base containing the REA objects and their relationships. The search key generator is the inference engine that draws inferences by tracing appropriate links in the semantic network. Using these two elements, the construction of a search key list is accomplished by identifying the primary search key(s) or triggers and constructing a search space (i.e., a contextual base) for the exploration of supplemental search keys by matching semantically related object(s) in the semantic network. In order not to draw the semantic relationships out too far, special rule like "constrained spreading activation" is applied. Using the key list, the search key generator forms text queries to retrieve relevant articles from the article database.

4.4. The Request module

Finally, the Administration module allows users to manage three databases: Web site URLs databases, financial articles database, and the semantic network database. This module, in particular importance, enables the researcher to refine the semantic network over times for better performance.

5. The use of the system

Credit managers and bankers often focus on the liquidity position by looking into a firm's short-term assets and liabilities, or its working capital condition. To analyze a firm's liquidity position, we use the Current Ratio which is calculated by dividing Current Assets value by Current Liabilities value. In order to gain insights into the company's liquidity position, we need to know if there were important activities, such as large bank loans or significant increase in inventory or changes in accounting reporting procedures that might affect its working capital, hence, the current ratio. With the help of the Retrieve module of the system, financial articles related to “liquidity position” in terms of current ratio concept are retrieved as a result of the search for all of the article containing words or terms that are related to the concept of “asset,” “liability,” and “liquidity”. The relevant keywords in the semantic network include “cash,” "inventory," “account receivable,” “debt,” and so on.

7. Recommendations and conclusions

Systems development is a key research methodology that interacts with other methodologies, such as theory building, experimentation, and observation. The advancement of intelligent text processing for Web resources often comes from new systems concepts, but systems must be developed first to test and measure the underlying concepts. The system described in this paper is a result of such a proof-of-concept approach.

Further research work will include substantial testing of the system for precision and recall measurements and the refinement of the semantic network for performance improvement.

References

[1] R. Bingi, D. Khazanchi, and S.B. Yadav, "Comparative Analysis of Knowledge Representation Schemes," *Information Processing & Management* (31:2), 1995, pp. 233-245.

[2] P.R. Cohen and R. Kjeldsen, "Information Retrieval by Constrained Spreading Activation in Semantic Networks," *Information Processing & Management* (32:4), 1987, pp. 255-268.

[3] E.L. Denna, J.O. Cherrington, D.P. Andros, and A.S. Hollander. "*Event-driven Business Solutions: Today's Revolution in Business and Information Technology*", Business One Irwin, Homewood, IL, 1993.

[4] W.E. McCarthy, "An Entity-relationship View of Accounting Models," *The Accounting Review* (54:4), October 1979, pp. 667-687.

[5] H.A. Simon and A. Newell, "Human Problem Solving: The State of the Theory in 1970," *American Psychologist*, February 1971, pp. 145-159.

[6] G.H. Sorter, "An 'Event' Approach to Basic Accounting Theory," *The Accounting Review* (44:1), January 1969, pp. 12-19.