

Business Stakeholder Analyzer: An Experiment of Classifying Stakeholders on the Web

Wingyan Chung

Department of Operations and Management Information Systems, Leavey School of Business, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053. E-mail: wchung@scu.edu

Hsinchun Chen

Artificial Intelligence Lab, Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721. E-mail: hchen@eller.arizona.edu

Edna Reid

Department of Library Science, Clarion University, Clarion, PA 16214. E-mail: ereid@clarion.edu

As the Web is used increasingly to share and disseminate information, business analysts and managers are challenged to understand stakeholder relationships. Traditional stakeholder theories and frameworks employ a manual approach to analysis and do not scale up to accommodate the rapid growth of the Web. Unfortunately, existing business intelligence (BI) tools lack analysis capability, and research on BI systems is sparse. This research proposes a framework for designing BI systems to identify and to classify stakeholders on the Web, incorporating human knowledge and machine-learned information from Web pages. Based on the framework, we have developed a prototype called Business Stakeholder Analyzer (BSA) that helps managers and analysts to identify and to classify their stakeholders on the Web. Results from our experiment involving algorithm comparison, feature comparison, and a user study showed that the system achieved better within-class accuracies in widespread stakeholder types such as partner/sponsor/supplier and media/reviewer, and was more efficient than human classification. The student and practitioner subjects in our user study strongly agreed that such a system would save analysts' time and help to identify and classify stakeholders. This research contributes to a better understanding of how to integrate information technology with stakeholder theory, and enriches the knowledge base of BI system design.

Introduction

The current networked business environment has greatly facilitated information sharing, electronic commerce, and

partner collaboration (Applegate, 2003). Businesses rely on the Internet to conduct a wide range of activities, including buying and selling products, analyzing business relationships, and researching development opportunities (Li & Du, 2003). To stay competitive, companies rely on business intelligence (BI) to monitor the operating environment, to identify potential risks, and to devise competitive strategies (Fleisher & Blenkhorn, 2003; Gilad, 2004; Prescott & Miller, 2001). BI can be obtained by analyzing a company's (internal) operational data (such as the financial statements; see Rasmussen, Goldy & Solli, 2002; and sales and transaction records; see Hurd & Nyberg, 2004) and by studying the company's (external) competitive environment and stakeholders (e.g., supply market; see Handfield, 2006; and various stakeholders' concerns; see Chung, 2008a; Freeman, 1984). For example, a business analyst can leverage on the BI gathered from various interested parties to formulate a strategic plan for his company. There exist many standard IT solutions for analyzing a company's internal data, including statistical software, online analytical processing, data warehousing, and data mining (Turban, Aronson, & Liang, 2005; Whitehorn & Whitehorn, 1999). In contrast, methods and technologies for studying a company's external environment and various stakeholders are less standardized due to the relatively unstructured nature of the data. The Web has provided a major channel for companies to share information about stakeholders. Important clues to knowledge about stakeholder identities and relationships are often expressed in textual content or annotated hyperlinks on Web pages.

As the Web is used increasingly to share and disseminate information, the problems arising from information overload and the interconnected nature of the Web make it difficult

Received October 1, 2007; revised July 17, 2008; accepted July 17, 2008

© 2008 ASIS&T • Published online 12 September 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20948

to obtain BI and to identify stakeholders. Information overload occurs when not all Web pages about stakeholders can be processed and utilized by a human user, leading to an inadequate understanding of the competitive environment (Bowman, Danzig, Manber, & Schwartz, 1994; Rogers & Agarwala-Rogers, 1975). The hyperlinked Web environment supports extensive interconnection among Web pages, enhancing communication among stakeholders while creating disorientation among Web users (Nielsen, 1990) and aggravating the information overload problem. Consequently, important yet complex stakeholder relationships are buried in voluminous interconnected Web pages. Business analysts may not be aware of many of a company's stakeholders, who may have current or future relationships. Knowledge is often hidden in interconnected Web resources, posing challenges to identifying and classifying various business stakeholders on the Web.

Researchers have developed theories and approaches for stakeholder management that help to devise strategies to understand and manage stakeholder relationships (e.g., Clarkson, 1995; Jawahar & McLaughlin, 2001). Although these theories and frameworks offer rich theoretical foundations for understanding stakeholder relationships, they are largely manually-driven and not scalable to accommodate the rapid growth and change of the Web. The proliferation of electronic commerce in recent years has increased the connection among business stakeholders on the Web, thus further complicating their relationships. So there is a need for better approaches to uncovering knowledge that may improve understanding of business stakeholder relationships. Unfortunately, existing BI tools lack the capability to support stakeholder analysis on the Web (Fuld, Singh, Rothwell, & Kim, 2003).

To address the aforementioned needs, this research proposes a framework for designing BI systems to identify and to classify business stakeholders on the Web. Grounded on stakeholder theory and using techniques in Web mining and Web-page classification, the framework consists of steps to collect and extract stakeholder information and to classify stakeholders automatically into various types, such as customer, partner, supplier, media, and government. Based on the framework, we have designed and developed a business stakeholder analysis system that helps managers and analysts to identify and to classify their stakeholders on the Web. The system incorporates both human knowledge and business Web site content to support stakeholder analysis effectively and efficiently. To understand the extent to which the system accomplished its intended purpose, we conducted an experiment with practitioner business users and student subjects to compare its performance level with a traditional manual stakeholder analysis method. We also experimented with different algorithms and Web-page features in the classification to study the performance using different system configurations. The questions addressed in this research areas follows:

- How can a framework for designing BI systems to identify and to classify stakeholders on the Web be developed?

- How can Web-page features be used in a prototype developed based on the framework to classify business stakeholders?
- How does the prototype perform in comparison with human judgment in terms of effectiveness, efficiency, and user perception in stakeholder classification?

Using the framework, we aimed to facilitate business stakeholder analysis on the Web by integrating information technology with stakeholder theory. At the company (i.e., microscopic) level, individual business stakeholders can be identified and classified. At the competitive environment (i.e., macroscopic) level, groups of business stakeholders can be formed to conduct further analysis.

Literature Review

The idea that organizations have stakeholders is well conceived in the literature (Freeman, 1984; Mitchell, Agle, & Wood, 1997). It has been suggested that an organization's survival and continuing prosperity depends on its ability to create sufficient wealth, value, or satisfaction for all primary stakeholder groups (Clarkson, 1995). Advances in information technology have enabled managers to serve and to understand its stakeholders better, primarily through more effective and efficient collection, storage, and analysis of information from the business environment. In recent years, a new class of information technology known as business intelligence (BI) systems emerged to support such tasks (Negash, 2004). In this section, we review literature related to stakeholder theories and approaches to define important concepts and to describe research developments in this field. Then we introduce the technologies of BI systems, focusing on environmental scanning and competitive intelligence analysis. We also review technologies in Web-page classification, an emerging technology having potential to support stakeholder classification and business analysis.

Stakeholder Theories and Approaches

The term *stakeholders* is defined as individuals or organizations who affect or are affected by the accomplishment of the firm's objectives (Freeman, 1984). Examples of stakeholders include customers, suppliers, government agencies, the general public, financial institutions, and trade associations. Table 1 summarizes stakeholder types considered in recent research. Freeman introduced the concept of stakeholder management in the strategic-management literature, arguing that firms should attend to the interests of all their stakeholders rather than just the shareholders. Researchers generally assume that managers of a firm are able to identify all its stakeholders and that an important responsibility of management is to attend to all stakeholders' needs (e.g., Donaldson & Preston, 1995). While the use of information technology has helped managers to collect and store information about stakeholders, the voluminous nature of this information challenges managers to understand stakeholders' needs and the competitive environment. Clarkson (1995) considered stakeholder data in the 1980s and does not

TABLE 1. Stakeholder types¹ considered in recent research.

Research	P	E	C	S	U	M	G	R	V	O	T	F	I	N	No.
Reid (2003)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					10
Jawahar & McLaughlin (2001)	✓	✓	✓	✓			✓			✓	✓	✓		✓	9
Elias & Cavana (2000)	✓	✓	✓	✓		✓	✓			✓		✓		✓	9
Agle, Mitchell, & Sonnefeld (1999)		✓	✓	✓	✓		✓							✓	5
Donaldson & Preston (1995)	✓	✓	✓	✓			✓				✓		✓	✓	8
Clarkson (1995)	✓	✓	✓	✓			✓								5
No.	5	6	6	6	1	2	6	1	1	3	2	2	1	4	-

¹P = Partners/suppliers, E = Employees/Unions, C = Customers, S = Shareholders/investors, U = Education/Research institutions, M = Media/Portals, G = Public/Government, R = Recruiters, V = Reviewers, O = Competitors, T = Trade associations, F = Financial institutions, I = Political groups, N = Special Interest Groups/Communities (Note that the class "Unknown" is not included here). No. = Column or row sum.

consider more complex relationships (such as global business networking; see Parkhe, Wasserman, & Ralston, 2006) in the e-commerce environment. Although it was confirmed by empirical results (Agle, Mitchell, & Sonnefeld, 1999), the theory developed by Mitchell et al. (1997) does not provide a practical system for processing voluminous information of today's stakeholders.

Advances in electronic commerce have transformed the landscape of the business environment. Stakeholders who previously could not affect a firm can now interact with the firm through the Internet. Existing stakeholder theories are limited in the way they accommodate new information technologies. For example, Jawahar and McLaughlin (2001) concluded that their descriptive stakeholder theory might be limited to traditional businesses only. Adapting stakeholder theory to the e-government domain, Flak and Rose (2005) noted that the impact of information technology on stakeholder management has not yet been explored, and recommend identifying and classifying stakeholders and improving descriptive stakeholder models to reflect a better understanding of the connection between technology and stakeholder relationships. These directions point to a need for integrating IT into traditional stakeholder theories and frameworks, which assume only a manual approach to stakeholder analysis (e.g., Elias & Cavana, 2000; Reid, 2003). These theories and frameworks may need to be augmented by Web-based, automatic approaches to environmental scanning, information collection, and stakeholder analysis. In particular, business intelligence, obtained from the business environment, is likely to help in stakeholder analysis, and automated tools have been developed to exploit BI.

Business Intelligence Systems

Business intelligence systems enable organizations to understand their internal and external environments through systematic acquisition, collation, analysis, interpretation, and exploitation of information (Cronin, 2000; Nolan, 1999). Advanced BI systems often rely on Web-mining techniques to extract information and discover patterns on the Web automatically (Chung, 2008a; Gregg, & Walczak, 2006; Kosala & Blockeel, 2000). Web content mining helps analysts to identify key terms used to describe business stakeholders. Web

structure mining facilitates understanding of how the macroscopic environment relates to certain Web sites or pages. Examples of Web-mining techniques include metasearching (Chen, Chau, & Zeng, 2002), Google's PageRank algorithm (Brin & Page, 1998), and the Hyperlink Induced Topic Search algorithm (Kleinberg, 1999). The external-link pages can be seen to mirror social communication phenomena, such as strategic or tactical referral behavior, and pragmatic or common semantic interest in particular sites on the Web (Ingwersen, 1998).

Previous research has developed automated systems to exploit Web content and link structure information for discovering business intelligence. Examples include Ong, Tan, Ng, Pan, and Li (2001) and Tan, Foo, and Hui (2002). Link analysis and metasearching have been used to support BI analysis (Chau, Shiu, Chan, & Chen, 2007). Other research tried to identify a company's noncustomer Web communities using a manual analysis of hyperlinks referencing from stakeholders to a company (Reid, 2003). Information including site traffic, in-link, content depth, and usage ranking has also been used to study stakeholders (Byrne, 2003). But the analysis provided was shallow because only simple statistics were studied. From previous research, we find Web content and structural information to be important for understanding BI and stakeholders. However, existing tools lack the analysis capability to provide such understanding (Fuld, Sawka, Carmichael, Kim, & Hynes, 2002), calling for a need to automate stakeholder identification and classification, a primary step of stakeholder analysis.

Web-Page Classification

Web-page classification is the process of assigning pages to predefined categories. Machine learning has been widely used to automate this process. Major approaches include *k*-nearest neighbor, neural network, support vector machines, and naïve Bayesian network (Chen & Chau, 2004). An important step in these approaches is to form the set of features to be used for classification.

Web-page textual content has been considered an important feature for classification. Kwon and Lee (2003) used a *k*-nearest neighbor approach to classifying selected Web pages that was extended to classification of Web sites.

Mladenec (1998) used a naïve Bayesian classifier and the Yahoo! Directory to automatically classify Web pages, each of which was represented by feature vectors containing n -gram (up to 5 words) with stop words removed, achieving low accuracies ranging from 25% to 50%. Sebastiani (2002) discusses extensively major machine-learning approaches to text categorization. Apart from textual features, structural features have been used in Web-page classification. Furnkranz (1999) employed the structure of an HTML representation and the structure of the Web to represent Web pages that were classified by a RIPPER learning algorithm, achieving average recall and precision of 78% and 87%. Glover, Tsioutsoulouklis, Lawrence, Pennock, and Flake (2002) used anchor texts and nearby words from pages that linked to the target page. Their data-driven feature-selection method led to high descriptive power, but the research did not explore multicategory classification, a typical requirement in stakeholder classification. Lee, Hui, Cheuk, and Fong (2002) applied neural networks to filtering pornographic Web content and compared Kohonen's SOM and Fuzzy ART neural networks in binary classification, where KSOM achieved a slightly higher accuracy than ART (92.8% vs. 87.9%), but KSOM took significantly more time in the training process. Other examples of Web-page classification can be found in Zhu, Yu, Chi, and Gong (2007) and Shen, et al. (2004). The WebKB project is an earlier effort in Web-page classification (<http://www.cs.cmu.edu/~WebKB>). With data collected before 1998, their manually tagged Web-page collection contains mainly university Web pages not suitable for business-intelligence analysis. A comprehensive review of Web-page classification algorithms and features can be found in Qi and Davison (in press).

In summary, previous research has used different features and feature-selection methods in Web-page classification. The features include (a) page textual content: full text, page title, headings; (b) link-related textual content: anchor text, extended anchor text, URL strings; and (c) page structural information: number of words, number of page out-links, inbound outlinks (i.e., links that point to its own company), outbound outlinks (i.e., links that point to external Web sites). The feature selection methods include (a) human judgment; (b) feature ratios and thresholding; and (c) use of a domain lexicon. When applied to Web-based business stakeholder analysis, Web-page classification helps to discover companies' interest groups on the Web and to enable companies to better understand the competitive environment. Surprisingly, this area has not been widely explored.

A Framework for Business Stakeholder Analysis on the Web

Electronic commerce has greatly facilitated transaction processing in recent years. However, increased sharing of information on the Web has added complexity to identifying and understanding business stakeholder relationships. Although previous research in stakeholder analysis provides rich theoretical background, conclusions drawn from

data collected before the mid-1990s may not reflect rapid developments in e-commerce. Extant stakeholder theories and frameworks were mainly developed for analyzing stakeholders of traditional businesses (Clarkson, 1995; Jawahar & McLaughlin, 2001). Their approaches are manually driven and do not scale up to accommodate the rapid growth and change of the Web. There is a need to integrate information technology into these theories and frameworks to support stakeholder analysis and intelligence gathering. Unfortunately, BI systems available nowadays lack stakeholder analysis capability, and research on BI is sparse (Negash, 2004). While various Web-page classification techniques have been developed, they have not been applied to business stakeholder classification. Therefore, a framework for designing BI systems to support identification and classification of stakeholders on the Web would be valuable to system developers, information systems researchers, and business practitioners. BI systems developed based on such a framework would help automate a significant part of intelligence gathering and stakeholder analysis.

Steps in the Framework

The rationale for the proposed framework (see Figure 1) is twofold. First, business stakeholders who have an interest in a company often have on their Web pages identifiable clues (e.g., textual descriptions, hyperlinks) that can be used to distinguish their stakeholder type. Second, Web content and structural information is important to understanding clues to stakeholder classification. Relying on such clues, we tried to automate stakeholder classification, an important step in stakeholder analysis.

The framework consists of three steps: intelligence gathering, tagging and feature selection, and automatic classification. Input to and output from the framework are, respectively, Web data and BI discovered after applying the steps. Each step allows human knowledge to guide the application of techniques (e.g., domain knowledge in data collection, manual tagging of Web pages, algorithm selection). The processed results include features and indexes extracted from Web pages and categories of stakeholders classified based on the extracted features. The arrows shown on the right of each step indicate that output of a step becomes input of the next step. As we move from bottom to top in these results, the degree of context and the difficulty of detecting noise in the results increase. The resulting stakeholder groups classified by the framework can be used to support business analysis and BI discovery. We explain the steps of the framework as follows.

Intelligence gathering. The purpose of this step is to gather relevant data from the Web to develop a collection of business Web pages. From these pages, we obtain several types of data: textual content (the text that can be seen on a Web browser), hyperlinks (embedded behind anchor text), and structural content (textual markup tags that indicate the types of content on the pages).

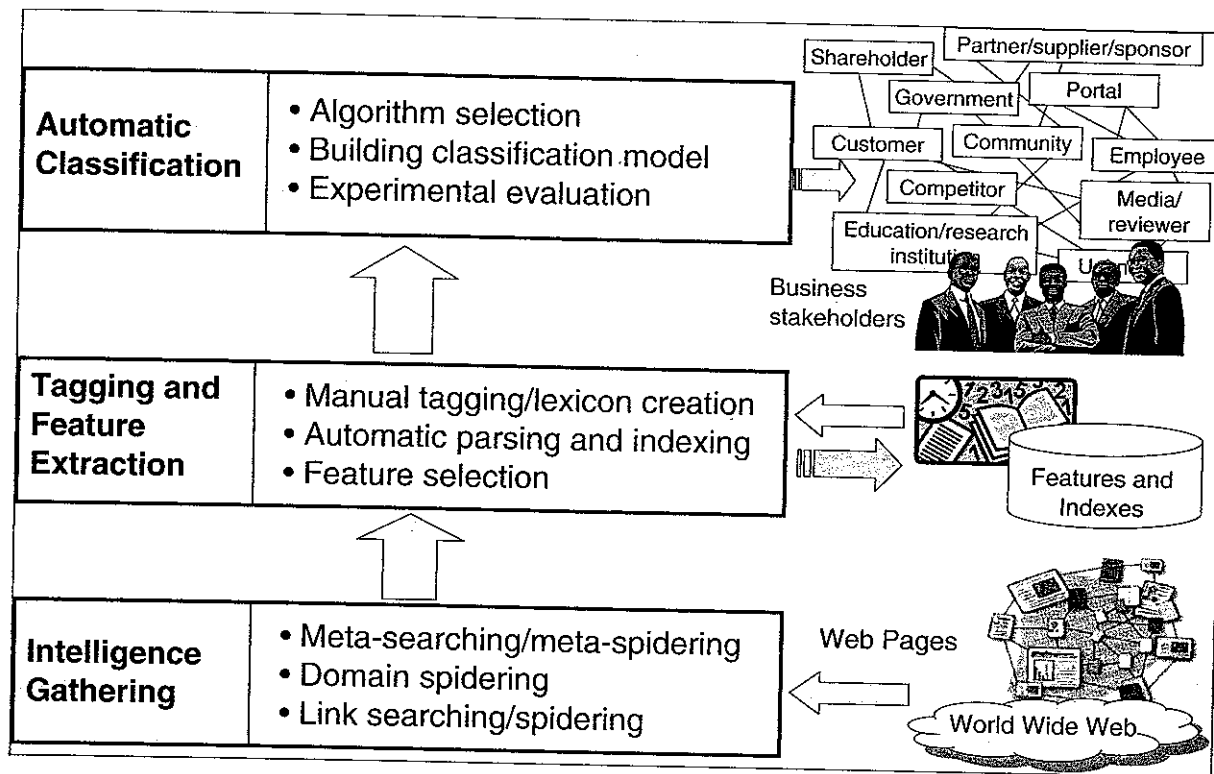


FIG. 1. A framework for business stakeholder analysis on the Web.

To collect these data, metasearching/metaspidering, domain spidering, and link searching/spidering are used. *Metasearching/metaspidering* uses keywords as inputs to search multiple Web search engines to collate a set of results (URL links) ranked among the top-ranked results in each engine. These keywords can be identified by human experts or by reviewing related literature. The process follows the links of the results and downloads appropriate Web pages for further processing. *Domain spidering* uses a set of seed URLs (provided by experts or identified in reputable sources) as starting pages. A crawler follows links in these pages to fetch pages automatically. Oftentimes, a breadth-first search strategy is used by the crawler because this generally provides good coverage of resources on the topic being studied. *Link searching/spidering* uses URL links as inputs to some search engines that support searching for Web pages containing these links in their content. For example, a user inputs the link "www.nytimes.com" to a search engine that returns other Web pages having hyperlinks referencing the www.nytimes.com Web site. Google performs such a search when the input link is preceded by a "link:" directive. Similar to metaspidering, the process of link-spidering follows the links of the results and downloads the Web pages. The result of this step is a collection of Web pages and documents that may contain much noisy data.

Tagging and feature extraction. This step aims to extract important features from Web pages and to tag a subset of pages representing different types of business stakeholders. *Manual tagging/lexicon creation* is a process of identifying

a subset of Web pages and assigning these pages manually to their respective types, where the types are often predefined using some taxonomies or classification schemes. Important keywords that distinguish Web pages into different types are identified and stored in a lexicon for assisting in automatic classification. *Automatic parsing* tries to obtain the structure of the content of the Web page from which features are extracted. For example, an HTML page uses tags to indicate such different types of Web content as headings, tables, and metadata. *Indexing* is the process of extracting terms (words or phrases) from textual documents and associating these terms with the documents. A list of stop words is typically used to remove non-semantic-bearing terms (e.g., *of, the, a*), which can be identified in the literature (e.g., van Rijsbergen, 1979). *Feature selection* extracts important textual, hyperlink, and structural information from Web pages to support classification, and can be done manually (with expert guidance) or automatically (using machine-learning algorithms). An example of a feature is a company name (e.g., Siebel, ClearForest) appearing on a Web page. The results of this step are features (e.g., terms and hyperlinks) and indexes (e.g., indicating which terms appear on which pages, showing the stakeholder relationship between a business and its partner). They provide more contextual information useful for automatic classification. Noise in data is reduced significantly from the previous step.

Automatic classification. The features and indexes extracted from the previous step can be further analyzed and processed to classify business stakeholders into their respective

types, thus facilitating understanding of the competitive environment. *Algorithm selection* identifies suitable methods to perform the classification task, often based on the criteria of classification accuracy and robustness of the methods. The *building classification model* is a process for the selected algorithms to learn automatically the association between the extracted features and the types to which the pages belong. The result of such learning is a classification model that can be used to predict the types of new pages. *Experimental evaluation* is a systematic procedure of assessing the performance of the automatic classification algorithms, often through comparison with such ground-truth alternatives as human classification and different algorithms and features. The measures used to judge the performance include accuracy, efficiency, and user ratings.

Business Stakeholder Analyzer

To demonstrate the feasibility and usability of the framework, we have developed a prototype called Business Stakeholder Analyzer (BSA) that supports automatic identification and classification of business stakeholders on the Web. The following describes our application of the steps in the framework to developing BSA.

Step 1: Intelligence gathering. We have collected Web pages of business stakeholders of the top 100 knowledge-management companies identified by the *Knowledge Management World Web* site (McKellar, 2003; <http://www.kmworld.com/>), a major Web portal providing news, publications, online resources, and solutions to more than 51,000 subscribers in the knowledge-management-systems market. To identify such stakeholders, we used the backlink search function of the Google search engine (<http://www.google.com/>) to search for Web pages having hyperlinks pointing to the companies' Web sites. This method has been successfully used to analyze the noncustomer online communities of a company (Reid, 2003). To illustrate the method, we can type "link:www.siebel.com" in Google's search box to find the Web pages pointing to Siebel's Web site (the host company). A relationship exists between Siebel and the results because the hyperlinks imply underlying stakeholder relationships with the company.

For each host company, we considered only the first 100 results returned from Google in order to limit the scope of analysis. We removed results that came from the same host company (i.e., self-links) and used only the first result, if more than one result came from the same Web site (by recognizing the domain name of the results' URLs). After filtering, we obtained 3,713 results in total. On average, we identified 37 stakeholders for each host company.

Among the stakeholders of the 100 host companies, we selected those of nine companies as training examples (see Table 2) for building a domain lexicon and for training the algorithms (discussed in the following sections). Sampled randomly from all the host companies, this subset of nine companies was therefore a representative sample of the 100

TABLE 2. Companies selected as training and testing examples.

Company	URL
Training examples	
Autonomy	http://www.autonomy.com
ClearForest	http://www.clearforest.com
Documentum	http://www.documentum.com
Fujitsu Software	http://www.fsc.fujitsu.com
Information Builders	http://www.informationbuilders.com
Plumtree	http://www.plumtree.com
SiteScape	http://www.sitescape.com
TheBrain	http://www.thebrain.com
West Group	http://west.thomson.com
Testing examples	
Applied Semantics	http://www.appliedsemantics.com
Computer Associates	http://www.cai.com
Dialog	http://www.dialog.com
Factiva	http://www.factiva.com
Intelliseek	http://www.intelliseek.com
Kamoon	http://www.kamoon.com
Siebel	http://www.siebel.com
Stratify	http://www.stratify.com
Tacit Knowledge Systems	http://www.tacit.com
WebMethods	http://www.webmethods.com

companies having close to 4,000 stakeholders. These 361 stakeholder pages of these nine companies were then automatically spidered, parsed, and indexed to extract textual terms.

Step 2: Tagging and feature extraction. An expert in business intelligence helped us to create a domain lexicon to include one-, two-, and three-word terms that were indicative of business stakeholder types. With over thirty years of information-systems experience, the expert holds doctoral and master's degrees in library science and a postgraduate certificate in management information systems, and was president of the Society of Competitive Intelligence Professionals in a developed Asian country. The lexicon-creation process involved generation of a list of stakeholder types and extraction of terms that are indicative of these stakeholder types, which are shown in Table 3, developed based on our review summarized in Table 1. The expert manually filtered out Web pages if (a) hyperlinks of host companies did not exist in the pages, (b) they contained too little text (fewer than 20 words), or (c) they mainly contained non-English content. This filtering resulted in 331 pages that were then used in the manual tagging and lexicon creation. The lexicon terms were extracted manually by the expert who read through all the Web pages of the nine companies' business stakeholders. For example, terms such as *news*, *newsletter*, and *news archives* are indicative of the media type. Terms such as *partner*, *alliances*, and *strategic partnership* are indicative of the partner/supplier/sponsor type. After analyzing all the pages, the expert extracted a total of 329 terms (67 one-word terms, 84 two-word terms, and 178 three-word terms) that would constitute our lexicon. The incorporation of expert

TABLE 3. Stakeholder types used in manual tagging of Web pages.

Group	Description	Stakeholder type	Number of training pages
Transactional (internal environment)	Actor that the enterprise interacts with and influences	Partner/supplier/sponsor	88
		Customer	13
		Employee	7
		Shareholder	3
Contextual (external environment)	Distance actor that the enterprise has no power or influence over	Government	1
		Competitor	1
		Community (special interest groups)	52
		Education/research institution	23
		Media/reviewer	77
		Portal creator/owner	59
Other	Unknown	Unknown	11

judgment into BSA development helped enrich the sources used for classification, thus addressing the call for better utilizing human knowledge in Web-page classification (Qi & Davison, in press).

The expert also manually classified each of the stakeholder pages of the nine companies into one of the 11 stakeholder types listed in Table 3. Each Web page was reviewed on the basis of terms and clues indicative of its type. If the texts or HTML tags on a Web page did not carry prespecified terms (those already in the domain lexicon) uniquely identifying a stakeholder type, generic terms were assigned by the expert on the basis of the manifest information of the Web page. The expert conducted an iterative and labor-intensive refinement process to ensure the consistency of outcomes, which were used to guide the process of automatic classification that used the manually classified Web pages as training examples.

Additional filtering of the stakeholder pages was done using a computer program and the aforementioned rules, resulting in 283 Web pages used for training and feature extraction. Two sets of features of business stakeholders' Web pages were selected to be input for automatic classification: structural content features and textual content features. Structural content features represent occurrences of lexicon terms in different parts of the Web page. Each structural content feature is a binary variable showing the occurrence (indicated by a 1) or absence (indicated by a 0) of a lexicon term in a part of a Web page. We have considered terms appearing in three parts of a page: page title, extended anchor text (the anchor text plus 50 words surrounding it), and page full text. These parts have been successfully used in previous research (Chung, Chen, & Nunamaker, 2005; Kwon & Lee, 2003; Lee et al., 2002) because they can reflect the importance of the terms. So there are in total 987 ($= 3 \times 329$) structural features. To identify such occurrences, an HTML parser automatically extracted all one-, two-, and three-word terms from the pages' full-text content. A list of 462 stop words was used to remove non-semantic-bearing words (e.g., *the, a, of, and*). Using HTML tags, the parser identified positions in which the terms appeared on the page.

Figure 2 shows an example of the HTML source code and a screen shot of the Web page of a business stakeholder of ClearForest, a company that provides content-management software. The title is "David Schatsky: Search and Discovery in the Post-Cold War Era." The extended anchor text includes "ClearForest" and terms surrounding it: "I just saw a demo by" and "a company that provides tools for analyzing unstructured textual information." The parser automatically checked the presence or absence of lexicon terms in the page title, extended anchor text, and full text.

Textual content features are the frequencies of occurrences of important one-, two-, and three-word terms appearing in the business stakeholder pages. By considering terms appearing in multiple categories of stakeholders, we modified the thresholding method used in Glover et al. (2002) to select important terms from a large number of extracted terms. Figure 3 shows the formulae and procedure used in the method. Terms with high feature ratios were selected as features for classification. Through the procedure, we could retain features that had high discriminating power among the stakeholder categories. Examples of such selected terms included *portals, companies, knowledge, coalition of the, portals research, building Web services, onlinetrade links to, and system design*. Features that rarely appeared in different categories were removed automatically. The rationale behind the procedure was to provide the algorithms with high-quality features as input, thereby enhancing the performance of classification. Because of its statistical nature, we believed that the selected features could help to differentiate the pages into stakeholder types. This automated method also reduced the need for labor-intensive human work.

Step 3: Automatic classification. Two machine-learning algorithms, feedforward/backpropagation neural network and support vector machines (Cristianini & Shawe-Taylor, 2000; Vapnik, 1995), were used to classify business stakeholder pages automatically into their respective stakeholder types. Neural network (NN), a computing system modeled after the human brain's meshlike network of interconnected neurons, has been shown to be robust in classification

```

<html><head>
<meta http-equiv="Content-Type"
content="text/html;
charset=iso-8859-1" />

<title>David Schatsky: Search
and Discovery in the Post-Cold
War Era</title>
...
<p>I just saw a demo by <a href
= "http://www.clearforest.com">
ClearForest, </a> a company
that provides tools for
analyzing unstructured textual
information. It's truly
amazing, and truly the search
tool for the post-Cold War era.
... </p>
</body></html>

```

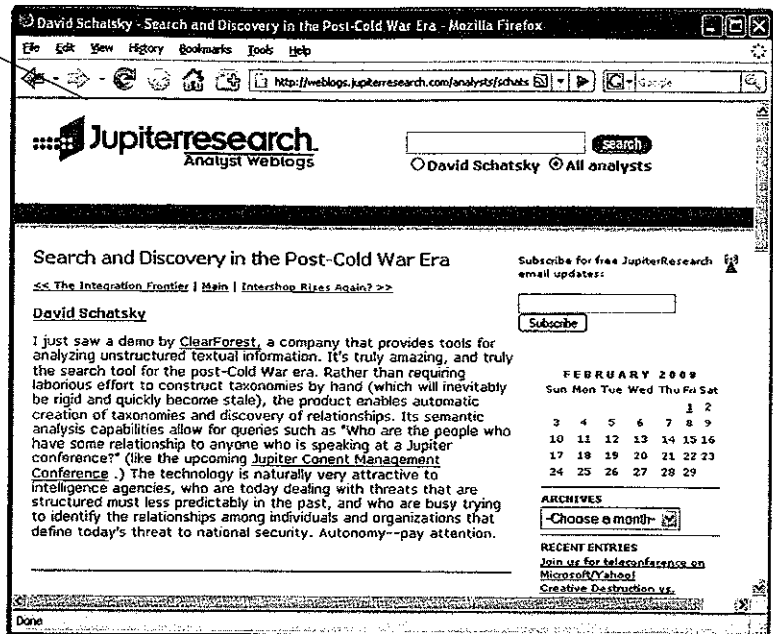


FIG. 2. A business stakeholder Web page of ClearForest.

Step 1: Suppose we have n classes of stakeholders and m features of Web pages. For each stakeholder class and a specific feature f_i , calculate the feature ratios (R_{f_i} and \tilde{R}_{f_i}) as follows:

$$R_{f_i} = \sum_{j=1}^n \frac{|C_{j,f_i}|}{|C_j|} \quad \text{and} \quad \tilde{R}_{f_i} = \sum_{j=1}^n \frac{|\tilde{C}_{j,f_i}|}{|\tilde{C}_j|}$$

where C_j = Web pages in class $j \in [1 \dots n]$
 C_{j,f_i} = Web pages in class j that contain feature f_i where $i \in [1 \dots m]$
 \tilde{C}_{j,f_i} = Web pages not in class j that contain feature f_i
 \tilde{C}_j = all the Web pages not belonging to class j

Step 2: Sort R_{f_i} and \tilde{R}_{f_i} . For each sorted list of R_{f_i} and \tilde{R}_{f_i} , select the top K features that have the highest values of the feature ratio, where K is approximately the number of structural-content features (In our testbed, K was equal to 1000, which was close to the number of structural-content features ($= 3 \times 329 = 987$), because there were 329 terms in our domain lexicon and we considered 3 positions (title, extended anchor text, and full text) where the terms would appear in Web pages). We try to assign equal importance between structural- and textual-content features because both types of content contribute equally to distinguishing the stakeholder classes. Then, two lists of features are obtained.

Step 3: Remove duplicating features appearing in the two lists.

FIG. 3. Formulae and procedure in the thresholding method.

and has wide applicability in different domains (Lippman, 1987) and Web-page filtering (Lee et al., 2002). Support vector machines (SVM), a machine-learning algorithm that tries to minimize structured risk in classification, has been successfully applied to text categorization (Joachims, 1998) and Web-page classification (Glover et al., 2002).

Each stakeholder page was represented as a feature vector containing 987 structural content features (binary variables indicating whether certain lexicon terms appeared in the page title, extended anchor text, and full text) and 1,297 textual content features (frequencies of occurrences of the selected features). These features were selected from about 200,000 words and phrases (see above for the selection procedure) that

were extracted from the stakeholder Web pages and from the many types of Web-page structural features that could be used in the classification. We used the features of the stakeholder pages of the nine selected companies to train the NN and SVM algorithms. The neural network is characterized by an input layer with 2,284 nodes (987 structural content features and 1,297 textual content features), a hidden layer with 1,148 nodes, and an output layer with 11 nodes (the 11 stakeholder classes). A single hidden layer was used in NN because it can model any complex system with desired accuracy (Patuwo, Hu, & Hung, 1993). To achieve high accuracy while avoiding overfitting, we chose the average of the number of input nodes and the number of output nodes to be the number of

hidden nodes (Rich & Knight, 1991). Also, the NN algorithm stopped when it had run for 100 epochs, which was chosen based on empirical testing. For the SVM implementation, we used the decomposition method for bound-constrained SVM formulation proposed in Hsu & Lin (2002b) to perform multiclass classification. The capacity constant (C) for running SVM was chosen to be 1, which was the default value in the SVM package developed by Hsu and Lin (2002a).

The model and weights obtained from the training were used to predict the types of business stakeholder pages of the 10 testing companies that were randomly chosen from the remaining host companies not used in the training dataset and are listed in Table 2. These stakeholder pages were manually classified by the aforementioned BI expert into their respective types.

Discussion. Combining domain-expert knowledge and a variety of information-collection, feature-selection, and classification techniques, the proposed framework tries to integrate information technology into stakeholder theories and frameworks to support stakeholder analysis and intelligence gathering. The contributions and novelty of the framework stem from three areas. First, the framework is to the best of our knowledge the first attempt to address systematically the need for stakeholder classification on the Web, thereby contributing to BI research and BI system design. The intelligence-gathering process, consisting of metasearching/metaspidering, domain spidering, and link searching/spidering, is a nontypical way of collecting data for stakeholder classification. Second, the feature-selection technique combines a statistical approach with the use of structural and textual features, thus capturing important information from voluminous Web data and allowing the input of valuable domain expertise. This new approach suitably accommodates the specific needs of stakeholder analysis on the Web. Third, the integration of intelligence collection and automatic Web-page processing in classifying stakeholder information is new. Instead of studying only text classification, where many techniques have been developed, we attempted to solve a new problem never addressed in previous literature. We believe that our collection of stakeholder Web pages is more current than many existing classification datasets (e.g., Reuters-21578, a commonly-used dataset for text classification, was assembled in 1987—over two decades ago), more relevant to stakeholder analysis, and more specific to Web analysis (as it contains not only textual but also structural information).

A Sample Scenario

To illustrate the benefits of using the proposed framework in business stakeholder analysis, we describe below a scenario of business-intelligence analysis and compare the results of using the framework versus using a traditional manual approach. A business analyst from Siebel (<http://www.siebel.com/>) would like to formulate a strategic

plan on gathering business intelligence and managing various interested parties of the company. He wants to know who has an interest in the company and what type of interest it is. A traditional approach to this stakeholder analysis would be to gather information manually from many sources, such as magazines, newspapers, government publications, and expert advice. Then the analyst manually digests all the collected information. He may use Web search engines to help, but he needs to formulate all search queries and choose search strategies based on his own knowledge. After getting the information, he has to manually classify the various parties into different stakeholder types to understand their interests in the company.

Alternatively, the analyst can use the proposed framework to obtain potential stakeholder Web pages automatically through Web-mining methods; use machine learning to read through the pages and generate analysis models; and automatically classify the stakeholders into different types. For example, from a hyperlink obtained by using the framework (<http://www.cic.com/partners/>), the analyst identifies that Communication Intelligence Corporation (CIC) partners with Siebel to integrate e-signature technology for sample delivery of regulated drugs. Based on the automatic classification results, the analyst learns that *CRM Daily* is a Web portal that has reported Siebel's Universal Application Network (<http://www.crm-daily.com/perl/story/20142.html>). Also, the analyst finds that Siebel is one of the over 400 original equipment manufacturers (OEM) customers of Hummingbird, a leading enterprise-software solution provider in the United States. These results would have been obtained less efficiently and less accurately had the analyst used a traditional manual approach. Moreover, knowing these relationships would enable analysts and managers of Siebel to plan better for addressing stakeholders' needs and to capitalize on the relationships. For example, Siebel's analyst can further evaluate the performance of Hummingbird's software solutions to see whether they should switch to another vendor or seek outsourcing from other countries. Also, the analyst can further study *CRM Daily* to find other Siebel competitors that may threaten Siebel's market-leading position. Similarly, the analyst can study CIC's Web site to learn about its other partners who may compete with or have potential future relationships with Siebel.

System Evaluation and Experimental Design

This section describes the methodology used to evaluate the performance of the prototype developed based on the framework. The evaluation objectives were to understand the extent to which the prototype assisted in automatic stakeholder classification and to gauge the performance levels of different classification methods used by different people (novices and practitioners). We explain below the experimental design, hypothesis testing, and experimental procedures.

Experimental Design

The experiment consisted of comparison of classification methods and a user evaluation study. In method comparison, we compared the performance levels of three classification methods: neural network (NN), support vector machine (SVM), and a baseline method. We created the baseline method by randomly assigning the stakeholders into different stakeholder types. This method served to indicate the gain in effectiveness from our proposed framework, thereby demonstrating the benefits of our framework and supporting a worst-case analysis deemed appropriate in evaluating any automatic classification (Chen, Zhou, Hu, & Yoo, 2004). Other examples of using baseline methods involving randomness can be found in Hisamitsu, Niwa, and Tsujii (2000) and McGuffin, Bryson, and Jones (2001). To compare the three methods (baseline, NN, SVM), stakeholder pages of the testing companies listed in Table 2 were classified with each method independently. Then we compared their performance levels, which were measured by the efficiency (time used, in minutes), overall accuracy, and within-class accuracy, defined as

$$\text{Overall accuracy} = \frac{1}{n} \sum_{i=1}^n \frac{\text{Number of correctly classified stakeholders in sample } i}{\text{Number of all classified stakeholders in sample } i}$$

where n = number of stakeholder samples used for testing

$$\text{Within-class accuracy (class } x) = \frac{\text{Number of stakeholders correctly classified as class } x}{\text{Number of all stakeholders belonging to class } x}$$

Because of the many classes involved in the classification, we measured the accuracy, which can reflect more clearly the classification performance than precision and recall, which could provide as many as 242 numbers (= 2 metrics × 11 classes × 11 classes)—difficult to use to evaluate the overall performance.

In the user evaluation study, we recruited two groups of human subjects to perform manual stakeholder classification (see Table 4 for their profiles). The first group consisted of 36 undergraduate students, aged about 19 and recruited

from an introductory IS course in a major university in the U.S. The second group consisted of 22 business practitioners from an MBA program in a U.S. university. On average, participants in this group were about 35 years of age and had 9.6 years of working experience in such positions as marketing manager, director of medical organization, school administrator, and manufacturing engineer. These two groups of subjects were chosen to study how novices and practitioners performed in stakeholder classification. Each subject was introduced to stakeholder analysis and was asked to use our prototype, “Business Stakeholder Analyzer (BSA),” to browse companies’ stakeholder lists.

For each subject group, we randomly selected three companies (Intelliseek, Siebel, and WebMethods) from the testing companies (see Table 2) to be the targets of analysis. Each subject was randomly assigned one of these three companies’ 10 stakeholder pages to perform classification. Because all the companies in our sample were well established in the KM market, we assumed no significant difference in the difficulty of classifying different companies’ stakeholder pages. Using a small random sample of the testing companies’ stakeholder pages enabled the subjects to finish classifying all assigned stakeholder pages within the limited experimental time frame (less than an hour), while ensuring that the stakeholder pages assigned to each subject properly represented the population of testing companies’ stakeholders and were statistically comparable to the pages used in machine classification. We also assumed that subjects in each group were not experts in stakeholder analysis, thus avoiding any unfair outcomes. Figure 4 shows screen shots of BSA and of the stakeholders of Siebel. The subject could find definitions of the stakeholders from BSA’s front page and was also provided with a paper copy of the stakeholder list. Their task was to classify the 10 stakeholders into their respective stakeholder types. Upon finishing the task, the subject filled in a poststudy questionnaire to provide demographic information and to rate their perception on various aspects of the framework on a seven-point Likert scale.

To study the extent to which the subjects agreed with each other in stakeholder classification, we measured the intersubject agreement ratio. This measure helped us to understand the reliability of the subjects’ classification performance. Reliability can be considered as the ratio of the true level of the measure to the entire measure (Trochim, 2001). As our measurement consisted of categories, we calculated the agreement ratio, a widely used estimate of reliability

TABLE 4. Subjects’ profiles.

Dimension	Student subjects	Practitioner subjects
Computer usage	15–20 hours per week on average	30 hours per week on average
Gender	18 males, 18 females	19 males, 3 females
Education	32 undergraduate students, 2 with associate’s degree, 2 with bachelor’s degree	4 subjects with associate’s degree, 16 with bachelor’s degree, 1 with master’s degree, 1 with doctoral degree
Age	31 subjects aged between 18 and 25; 3 subjects aged between 26 and 30; 2 subjects aged between 31 and 35	5 subjects aged between 18 and 25; 9 subjects aged between 26 and 30; 4 subjects aged between 31 and 35; 2 subjects aged between 36 and 40; 2 subjects aged between 51 and 60

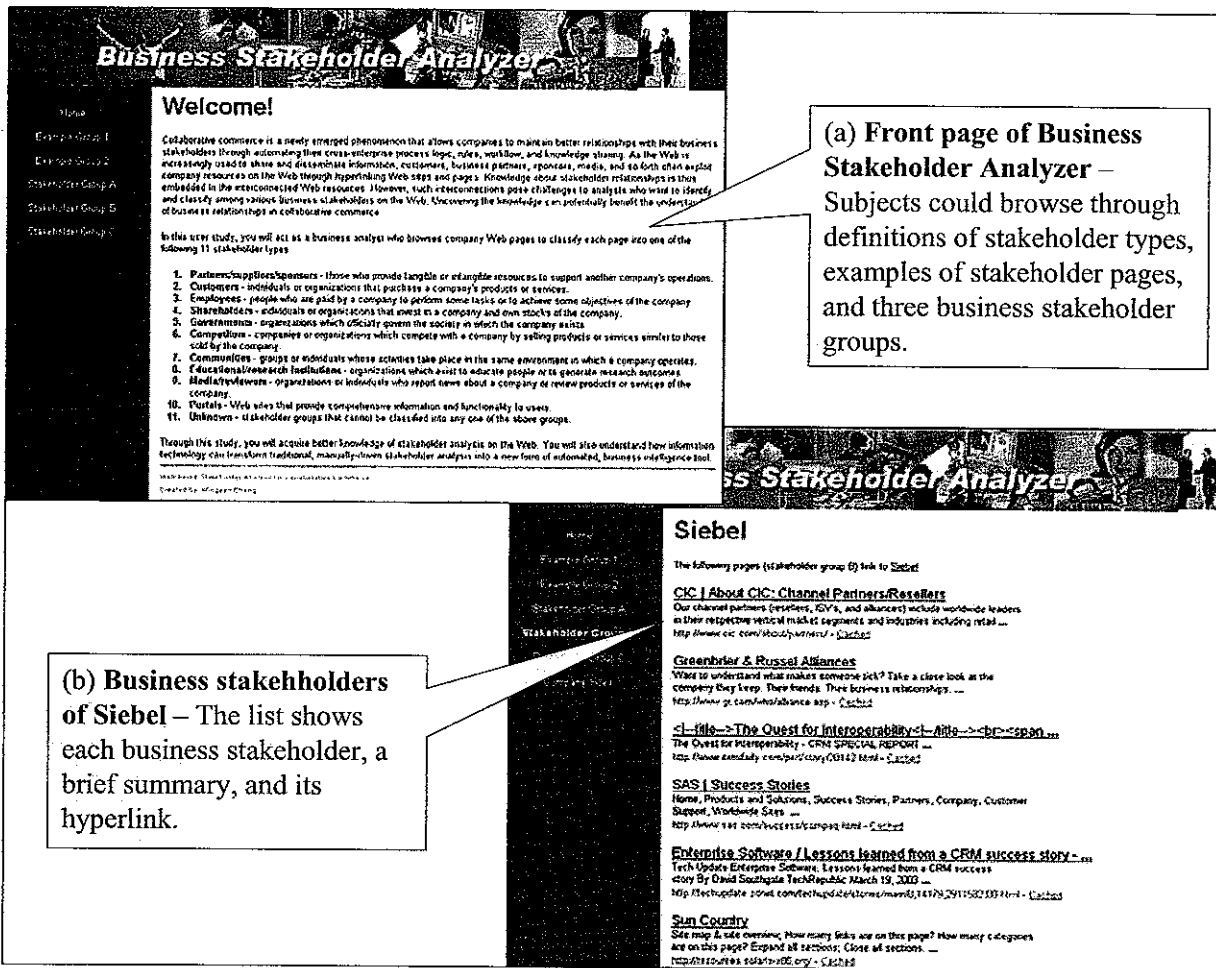


FIG. 4. Business Stakeholder Analyzer.

(Perreault & Leigh, 1989), between pairs of subjects to estimate interrater reliability, as shown in the following formula:

$$\text{Agreement ratio } (P_a) = \frac{\text{Number of agreements between a pair of subjects}}{\text{Total number of judgments compared between the subjects}}$$

For each possible pairwise combination of subjects among those assigned to the same selected company, we calculated the agreement ratio (P_a). The overall reliability was estimated by averaging the ratios (weighted by the number of all possible pairwise combinations among subjects assigned to a company) across the three selected companies in a group of subjects. For the undergraduate student subjects, 12 subjects were assigned to each of the three selected companies. For the practitioner subjects, 6 subjects were assigned to Intel-liseek, 8 subjects were assigned to Siebel, and 8 subjects were assigned to WebMethods.

Hypotheses and Experimental Procedures

Table 5 shows the two groups of 10 hypotheses tested in this study. The first group (H1–H2) hypothesized that the

two algorithms (NN and SVM) would perform better than the baseline method because both algorithms incorporated human knowledge and machine-learning capability into the classification, thereby adding value to business stakeholder analysis. To test the hypotheses, we created 30 sets of stakeholder pages by a random selection of stakeholder pages of the testing companies. Each set consists of 5 stakeholder pages on which the classification methods made predictions. The 30 sets of pages served as different classification scenarios to test the methods' performances.

The second group (H3–H4) hypothesized that human judgment in stakeholder classification would achieve effectiveness similar to that of automatic classification, but that the former is less efficient. The rationale was that both classification algorithms and human analysts could make use of the Web page's textual and structural content in classifying stakeholders. However, humans would require more time to understand and process the information. In all four hypotheses, effectiveness was measured by the overall classification accuracy defined above. Efficiency was measured by the time used in classification. To maintain a fair comparison, we did not consider the time used in training (because human subjects took a significantly longer time in acquiring classification knowledge than machine-learning algorithms,

TABLE 5. Hypotheses tested in this study.

<i>Method comparison</i>	
H1a-c:	NN achieves higher effectiveness than the baseline method (a) when structural content features are used; (b) when textual content features are used; (c) when both features are used.
H2a-c:	SVM achieves higher effectiveness than the baseline method (a) when structural content features are used; (b) when textual content features are used; (c) when both features are used.
<i>Comparing with human judgment</i>	
H3a-b:	Human judgment in stakeholder classification (a) achieves similar effectiveness to using NN (with both structural and textual content features); (b) is less efficient than using NN (with both structural and textual content features).
H4a-b:	Human judgment in stakeholder classification (a) achieves similar effectiveness to using SVM (with both structural and textual content features); (b) is less efficient than using SVM (with both structural and textual content features).

TABLE 6. Results of hypothesis testing.

Hypothesis ¹	Mean		SD		p-value	Supported?
	NN		Baseline			
H1a	0.50		0.25	0.08	0.00	Yes
H1b	0.31		0.21	0.08	0.00	Yes
H1c	0.19		0.16	0.08	0.01	Yes
	SVM		Baseline			
H2a	0.25		0.23	0.08	0.02	Yes
H2b	0.43		0.24	0.08	0.00	Yes
H2c	0.44		0.24	0.08	0.00	Yes
	Human		NN (in H3)/SVM (in H4)			
	S ³	P ³	S	P		
H3a	0.56	0.53	0.20	0.22	0.33	0.10
H3b ²	21.68	22.31	7.81	6.91	0.33	0.00
H4a	0.56	0.53	0.20	0.22	0.43	0.10
H4b ²	21.68	22.31	7.81	6.91	0.017	0.00

¹Effectiveness, ranging from 0 to 1, was measured by the overall accuracy.
²For hypotheses H3b and H4b, efficiency was measured by the time used (in minutes).
³"S" represents student subjects and "P" represents practitioner subjects.

which spent on average a few seconds to less than a minute on training).

Experimental Results and Discussion

This section reports and discusses the findings of our study. Table 4 summarizes subjects' profiles Table 6 provides the results of hypothesis testing. Table 7 lists the within-class accuracies achieved by different methods (NN, SVM, baseline, and human judgment). We also discuss subjects' comments, feedback, and intersubject classification reliability.

Effectiveness of BSA

The results of testing hypotheses H1a-c and H2a-c were all positive, confirming our belief that the use of the stakeholder classification framework would outperform the baseline method significantly. The framework has integrated human knowledge with machine-learned information related to stakeholder types. Automatic Web-page

classification of business stakeholder pages could alleviate information overload and make analysis more effective. The experimental findings showed that the performance of BSA was significantly better than the baseline method. We therefore conclude that it is promising to apply the framework to automating business stakeholder analysis on the Web.

To our surprise, we found that different feature sets yielded different performances of the algorithms. Structural content features enabled NN to achieve significantly better effectiveness than textual content features. On the other hand, textual content features enabled SVM to achieve significantly better effectiveness than structural content features. Furthermore, a combination of the two feature sets made NN less effective than using any one set alone. Future research can explore this issue further by studying the effect of features and the nature of algorithms.

Comparing BSA with Human Judgment

Overall accuracy. H3a and H4a were the only two hypotheses (out of 10) not confirmed, indicating that humans were

TABLE 7. Within-class accuracies achieved by different methods.

Stakeholder type	Frequency of occurrences	NN			SVM			B ²	Frequency of occurrences	S ²	p ²
		Structural	Textual	Combined	Structural	Textual	Combined				
Partners/suppliers/sponsors	37	0.62	0.86	0.70	0.97	0.97	0.97	0.08	156	0.65	0.62
Customers	4	0	0	0.25	0	0	0	0	24	0.42	0.21
Employees	3	0	0	0	0	0	0	0	0	-	-
Shareholders	2	0	0	0	0	0	0	0	0	-	-
Government	1	0	0	0	0	0	0	0	0	-	-
Competitors ¹	0	0	0	0	0	0	0	0	0	-	-
Communities	15	0	0	0	0	0	0	0	0	-	-
Education/Research institutions	6	0	0	0.17	0	0.27	0.27	0.13	0	-	-
Media/Reviewer	51	0.69	0.22	0	0	0.33	0.33	0.04	108	0.56	0.52
Portal	23	0.52	0	0	0	0.22	0.22	0.13	60	0.46	0.34
Unknown	1	0	0	0	0	0	0	0	0	-	-
Average	13	0.17	0.10	0.10	0.09	0.16	0.16	0.05	32.7	0.45	0.40
Overall accuracy	-	0.50	0.31	0.19	0.25	0.43	0.44	0.08	-	0.56	0.53

¹No stakeholder belonging to the type "competitor" was found.

²"B" stands for baseline method, "S" stands for undergraduate student subjects, and "P" stands for practitioner subjects.

in general more effective in the classification. We found that humans (students and practitioners) achieved the highest overall accuracy (0.56 and 0.53 respectively) among the sampled stakeholder types, while the best automatic method achieved a 0.5 overall accuracy. We believe that humans may rely on more clues in performing classification (e.g., contextual factors, graphics and other media on Web pages, and combinations of these). By using their experience in Internet browsing and searching and their domain knowledge, humans also could narrow down the potential types to which a stakeholder might belong. In contrast, the algorithms lacked such rich experience to aid classification.

Within-class accuracy. Taking a closer look at the within-class accuracies, we found that both NN and SVM outperformed humans in classifying some widespread stakeholder types. The within-class accuracies achieved by different methods are shown in Table 7, in which we also show the performance of the two algorithms when different types of features were used (textual, structural, and combined). SVM (using textual or combined features) achieved the best within-class accuracies for the partner/sponsor/supplier (0.97) and community (0.27) types. NN (structural) achieved the best within-class accuracies for the media/reviewer (0.69) and portal (0.52) types. These stakeholder types were among the most frequently occurring. In contrast, humans achieved lower within-class accuracies in all these types. Nevertheless, humans achieved the best within-class accuracies for the customer type (0.42, by student subjects) and educational/research institution type (0.33, by professional subjects), probably because humans could rely on their experiences as customers and were familiar with educational/research Web sites. Therefore, we conclude that humans achieved significantly higher overall accuracy than the two algorithms, but performed less well

than the algorithms in classifying stakeholder types that occur frequently in the dataset. This result indicates that the proposed framework can yield favorable within-class accuracies in classes where sufficient stakeholder instances exist. As a future direction, we plan to increase the size of the dataset to enhance the within-class accuracies in all the classes.

Although both the overall accuracies and within-class accuracies achieved by humans and algorithms were below 100%, these accuracies were significantly higher those achieved by the baseline method, indicating the actual gains from using the framework.

Hypotheses H3b and H3d were supported, demonstrating the high efficiency of using BSA to facilitate stakeholder analysis. The subjects took an average of 22 minutes to finish the task. Their times to completion varied a lot, ranging from the longest time of 42 minutes to the shortest of 11 minutes. In contrast, the machine-learning algorithms took from a few seconds to less than a minute to finish the classification. Such encouraging results led us to conclude that the framework could significantly augment human work. We believe that the framework has the potential to assist human analysts to automate part of their stakeholder-analysis work, thus allowing them to focus more on other value-added tasks.

As the neural-network algorithm achieved lower accuracies when both textual and structural features were used than when only one of these two types of features was used, it would be an interesting future study to identify the optimal number and types of features to use in the algorithm to achieve the best performance.

Intersubject Agreement

Subjects agreed with each other in approximately 40% of their judgments, even though each subject had to

independently choose from as many as 11 stakeholder types and to assign only one of these types to each stakeholder page. This agreement level is high considering their lack of experience and the multitude of stakeholder pages they had to classify within a short time. The student and practitioner subjects achieved similar average agreement ratios (40.15% and 40.14% respectively), indicating that their judgments served as a good benchmark to compare against BSA.

Users' Subjective Comments

Strong preferences toward an automatic approach to business stakeholder analysis were found in the user study. The proposed framework was perceived to be necessary to alleviate information overload on the Web (student rating = 2.28, practitioner rating = 2.00, where 1 = *Strongly agree*, and 7 = *Strongly disagree*) and to help business analysts identify and classify business relationships (student rating = 2.03, practitioner rating = 1.91). The subjects showed an overwhelming agreement on the statement that the framework would save the time of business analysts (student rating = 1.64, practitioner rating = 1.59). None of the subjects gave a rating of 6 or 7 (*Strongly disagree*) to any of the three statements. Many subjects provided favorable comments. The student subjects' comments include: "It would be very helpful!" and "I want to use it." A practitioner subject said, "Automation of this process is definitely beneficial to business analysts. Time is money and this process would save a lot of it." Another practitioner subject agreed that the framework "uses fewer resources and allows an analyst to get more details using the Web." One subject said, "It would allow for a better perception of a company's product since blogs are an important source of public relations (information)." From these results, we conclude that the framework was perceived very favorably as helping business analysts identify and classify stakeholders. It also confirmed our belief that the framework can support business stakeholder analysis.

Conclusions and Future Directions

As the Web is used increasingly to share and disseminate company and industry information, understanding stakeholder relationships has become an important area of business analysis. However, the large number of stakeholder classes and their voluminous information on the Web hinder this understanding. In this article, we have proposed a framework for designing BI systems to identify and classify stakeholders on the Web. Human expert knowledge and machine-learned information about business stakeholders have been integrated to enable effective and efficient analysis. Based on the framework, we have developed a business stakeholder analysis prototype (BSA) that supports stakeholder analysis on the Web. Results of our experiment involving method comparison and user evaluation showed that BSA significantly outperformed a baseline method and achieved the highest

within-class accuracies in classifying frequently appearing stakeholder types, while humans achieved the highest overall accuracy. Subjects in our user study strongly agreed that the framework would save business analysts' time and help in stakeholder analysis. There is a strong potential to use the framework to augment traditional stakeholder classification, as shown in the scenario described above. This research thus contributes to developing a useful framework for designing BI systems to identify and classify stakeholders on the Web, providing an example of integrating information technology with traditional stakeholder theories, and enriching the knowledge base of BI research and system design. The algorithm for stakeholder feature selection and its application are new. The use of expert domain knowledge and Web-page anchor text in BSA development directly addresses recommendations for future work in Web classification (Qi & Davison, in press).

This research has several limitations. Because some stakeholder types do not appear frequently in our training data (e.g., competitors, government, and shareholders), the classification algorithms might not have modeled these types accurately. The results would have been better if more training data with a wider distribution of stakeholder types were available. Also, the sizes of training and testing data prevented us from testing our framework in a more comprehensive manner. The classification accuracy would have been improved if a larger dataset were used. The use of Google to identify stakeholder Web pages might have limited our scope of data collection. On the other hand, despite expert participation in lexicon creation and Web-page tagging, the knowledge used for stakeholder classification was still limited and could have been better acquired had more experts participated in triangulating their views to improve the quality of the lexicon and tagging.

As classification is a beginning step in business stakeholder analysis, a promising future direction is to automate the next steps of such analysis. With more expert participation and more Web-page data, type-specific stakeholder analysis can be performed. For example, partner relationships are often important in developing business strategies. Gaining more specific knowledge about such relationships through automatic approaches is expected to help. Other potential sources for intelligence gathering include blogs, Web sites that support user interaction, and social-networking sites. In addition, stakeholder relationships form patterns over time. Tracing such patterns with automatic techniques such as visualization is likely to uncover knowledge about the competitive environment. Automating part of the lexicon-building process can help save the time and effort put into building it manually. Another future direction is automating cross-regional business stakeholder analysis and domain-specific stakeholder analysis. Multinational business partnerships and cooperation in specific domains can be analyzed through explicit information posted on the Web, which is used increasingly by non-English speakers around the world (Chung, 2008b). Related stakeholder theories and human-computer interaction issues can be explored.

Acknowledgments

We thank the subjects and expert who participated in the experiment. We also thank the editor and the anonymous reviewers for their comments and suggestions.

References

- Agle, B.R., Mitchell, R.K., & Sonnefeld, J.A. (1999). Who matters to CEOs? An investigation of stakeholders' attributes and salience, corporate performance, and CEO values. *Academy of Management Journal*, 42, 507-525.
- Applegate, L.M. (2003, October). Building businesses in a networked economy. Paper presented at Management Information Systems Fall Conference on Managing Information Technologies in Networked Organizations, Tucson, AZ.
- Bowman, C.M., Danzig, P.B., Manber, U., & Schwartz, F. (1994). Scalable Internet resource discovery: Research problems and approaches. *Communications of the ACM*, 37(8), 98-107.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the 7th International World Wide Web Conference, Article FPI 1. Retrieved September 2, 2008, from <http://www7.scu.edu.au/00/index.htm>
- Byrne, J. (2003). Answering the questions of the universe: Who am I and how do I fit in? Paper presented at Ragan Annual PR Conference, Chicago, IL.
- Chau, M., Shiu, B., Chan, I., & Chen, H. (2007). Redips: Backlink search and analysis on the Web for business intelligence analysis. *Journal of the American Society for Information Science and Technology*, 58, 351-365.
- Chen, H., & Chau, M. (2004). Web mining: Machine learning for Web applications. In M.E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST)* (pp. 289-329). Medford, NJ: Information Today.
- Chen, H., Chau, M., & Zeng, D. (2002). CI Spider: A tool for competitive intelligence on the Web. *Decision Support Systems*, 34, 1-17.
- Chen, H., Zhou, H.-X., Hu, X., & Yoo, I. (2004). Classification comparison of prediction of solvent accessibility from protein sequences. In Proceedings of the Second Asia-Pacific Bioinformatics Conference (APBC2004) (pp. 333-338), Dunedin, New Zealand: Australian Computer Society.
- Chung, W. (2008a). Visualizing E-Business stakeholders on the Web: A methodology and experimental results. *International Journal of Electronic Business*, 6, 25-46.
- Chung, W. (2008b). Web searching in a multilingual world. *Communications of the ACM*, 51(5), 32-40.
- Chung, W., Chen, H., & Nunamaker, J.F. (2005). A visual framework for knowledge discovery on the Web: An empirical study on business intelligence exploration. *Journal of Management Information Systems*, 21(4), 57-84.
- Clarkson, M.B.E. (1995). A stakeholder framework for analyzing and evaluating corporate social performance. *Academy of Management Review*, 20, 92-117.
- Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge, UK: Cambridge University Press.
- Cronin, B. (2000). Strategic intelligence and networked business. *Journal of Information Science*, 26, 133-138.
- Donaldson, T., & Preston, L.E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review*, 20, 65-91.
- Elias, A.A., & Cavana, R.Y. (2000). Stakeholder analysis for systems thinking and modeling. Paper presented at the 35th Annual Conference of the Operational Research Society of New Zealand, Wellington, New Zealand.
- Flak, L.S., & Rose, J. (2005). Stakeholder governance: adapting stakeholder theory to e-government. *Communications of the Association for Information Systems*, 16, 642-664.
- Fleisher, C.S., & Blenkhorn, D.L. (Eds.) (2003). *Controversies in competitive intelligence: the enduring issues*. Westport, CT: Praeger.
- Freeman, E. (1984). *Strategic management: A stakeholder approach*. Marshfield, MA: Pitman.
- Fuld, L.M., Sawka, K., Carmichael, J., Kim, J., & Hynes, K. (2002). *Intelligence Software Report™ 2002*. Cambridge, MA: Fuld & Company.
- Fuld, L.M., Singh, A., Rothwell, K., & Kim, J. (2003). *Intelligence Software Report™ 2003: Leveraging the Web*. Cambridge, MA: Fuld & Company.
- Furnkranz, J. (1999). Exploiting structural information for text classification on the WWW. In Proceedings of the Third Symposium on Intelligent Data Analysis (pp. 487-497). Amsterdam: Springer-Verlag.
- Gilad, B. (2004). *Early warning: Using competitive intelligence to anticipate market shifts, control risk, and create powerful strategies*. New York: American Management Association.
- Glover, E.J., Tsioutsoulidis, K., Lawrence, S., Pennock, D.M., & Flake, G.W. (2002). Using Web structure for classifying and describing Web pages. In Proceedings of the 11th International World Wide Web Conference (pp. 562-569). New York: ACM.
- Gregg, D.G., & Walczak, S. (2006). Adaptive Web information extraction. *Communications of the ACM*, 49(5), 78-84.
- Handfield, R. (2006). Supply market intelligence: A managerial handbook for building sourcing strategies (p. 637). Boca Raton, FL: Auerbach.
- Hisamitsu, T., Niwa, Y., & Tsujii, J.-I. (2000). A method of measuring term representativeness: Baseline method using co-occurrence distribution. In Proceedings of the 18th Conference on Computational Linguistics (pp. 320-326). Morristown, NJ: Association for Computational Linguistics.
- Hsu, C.W., & Lin, C.J. (2002a). A comparison on methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13, 415-425.
- Hsu, C.W., & Lin, C.J. (2002b). A simple decomposition method for support vector machines. *Machine Learning*, 46(1-3), 291-314.
- Hurd, M., & Nyberg, L. (2004). *The value factor: How global leaders use information for growth and competitive advantage* (p. 132). Princeton, NJ: Bloomberg Press.
- Ingwersen, P. (1998). The calculation of Web impact factors. *Journal of Documentation*, 54, 236-243.
- Jawahar, I.M., & McLaughlin, G.L. (2001). Toward a descriptive stakeholder theory: an organizational life cycle approach. *Academy of Management Review*, 26, 397-414.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning (pp. 137-142). Chemnitz, Germany: Springer-Verlag.
- Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association of Computing Machinery*, 46, 604-632.
- Kosala, R., & Blockeel, H. (2000). Web mining research: a survey. *ACM SIGKDD Explorations*, 2(1), 1-15.
- Kwon, O.-W., & Lee, J.-H. (2003). Text categorization based on k-nearest neighbor approach for Web site classification. *Information Processing & Management*, 39, 25-44.
- Lee, P.Y., Hui, S.C., Cheuk, A., & Fong, M. (2002). Neural networks for Web content filtering. *IEEE Intelligent Systems*, 17(5), 48-57.
- Li, E.Y., & Du, T.C. (2003). Emerging issues in collaborative commerce: Call for papers. *Decision Support Systems*, 35, 257-258.
- Lippman, R.P. (1987). Introduction to computing with neural networks. *IEEE ASSP Magazine*, 4(2), 4-22.
- McGuffin, L.J., Bryson, K., & Jones, D.T. (2001). What are the baselines for protein fold recognition? *Bioinformatics*, 17, 63-72.
- McKellar, H. (2003). KMWorld's 100 companies that matter in knowledge management 2003, KM World. [Online]. Retrieved September 2, 2008, from <http://www.kmworld.com/Articles/ReadArticle.aspx?ArticleID=41029>
- Mitchell, R.K., Agle, B.R., & Wood, D.J. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of Management Review*, 22, 853-886.
- Mladenic, D. (1998). Turning Yahoo into an automatic Web page classifier. In Proceedings of the 13th European Conference on Artificial Intelligence (pp. 473-474). Brighton, UK, John Wiley & Sons.

- Negash, S. (2004). Business intelligence. *Communications of the Association for Information Systems*, 13, 177-195.
- Nielsen, J. (1990). The art of navigating through hypertext. *Communications of the ACM*, 33(3), 296-310.
- Nolan, J. (1999). *Confidential: Uncover your competitor's secrets legally and quickly and protect your own*. New York: Harper Business.
- Ong, H.-L., Tan, A.-H., Ng, J., Pan, H., & Li, Q.-X. (2001). FOCT: Flexible Organizer for Competitive Intelligence. In *Proceedings of the 10th International Conference on Information and Knowledge Management* (pp. 523-525). New York: ACM.
- Parkhe, A., Wasserman, S., & Rafton, D.A. (2006). New frontiers in network theory development. *Academy of Management Review*, 31, 560-568.
- Patuwo, E., Hu, M.S., & Hung, M.S. (1993). Two-group classification using neural networks. *Decision Sciences*, 24, 825-845.
- Perreault, W.D., & Leigh, L.E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, 26(2), 135-148.
- Prescott, J.E., & Miller, S.H. (Eds.) (2001). *Proven strategies in competitive intelligence: Lessons from the trenches*. New York: Wiley.
- Qi, X., & Davison, B.D. (in press). Web page classification: Features and algorithms. *ACM Computing Surveys*.
- Rasmussen, N., Goldy, P.S., & Solli, P.O. (2002). *Financial business intelligence: Trends, technology, software selection, and implementation* (p. 283). New York: Wiley.
- Reid, E.O.F. (2003). Identifying a company's noncustomer online communities: a proto-typology. In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS-36)* (p. 215). IEEE Computer Society.
- Rich, E., & Knight, K. (1991). Learning in neural networks. In *Artificial Intelligence* (2nd ed., pp. 500-507). New York: McGraw-Hill.
- Rogers, E.M., & Agarwala-Rogers, R. (1975). Organizational communication. In G.L. Hanneman & W.J. McEwen (Eds.), *Communication Behavior* (pp. 218-236). Reading, MA: Addison-Wesley.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47.
- Shen, D., Chen, Z., Yang, Q., Zeng, H.-J., Zhang, B., Lu, Y., & Ma, W.-Y. (2004). Web-page classification through summarization. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 242-249). Sheffield, UK: ACM.
- Tan, B., Foo, S., & Hui, S.C. (2002). Web information monitoring for competitive intelligence. *Cybernetics & Systems*, 33, 225-251.
- Trochim, W. (2001). *The research methods knowledge base* (2nd ed.) Mason, OH: Atomic Dog.
- Turban, E., Aronson, J.E., & Liang, T.-P. (2005). *Decision support systems and intelligent systems* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- van Rijsbergen, C.J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths.
- Vapnik, V.N. (1995). *The nature of statistical learning theory*. New York: Springer.
- Whitehorn, M., & Whitehorn, M. (1999). *Business intelligence: The IBM solution—Data warehousing and OLAP* (p. 289). London: Springer.
- Zhu, S., Yu, K., Chi, Y., & Gong, Y. (2007). Combining content and link for classification using matrix factorization. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 487-494). Amsterdam: ACM.