Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers

TIANJUN FU, University of Arizona
AHMED ABBASI, University of Virginia
DANIEL ZENG, Institute of Automation, Chinese Academy of Sciences, and University of Arizona
HSINCHUN CHEN, University of Arizona

Despite the increased prevalence of sentiment-related information on the Web, there has been limited work on focused crawlers capable of effectively collecting not only topic-relevant but also sentiment-relevant content. In this article, we propose a novel focused crawler that incorporates topic and sentiment information as well as a graph-based tunneling mechanism for enhanced collection of opinion-rich Web content regarding a particular topic. The graph-based sentiment (GBS) crawler uses a text classifier that employs both topic and sentiment categorization modules to assess the relevance of candidate pages. This information is also used to label nodes in web graphs that are employed by the tunneling mechanism to improve collection recall. Experimental results on two test beds revealed that GBS was able to provide better precision and recall than seven comparison crawlers. Moreover, GBS was able to collect a large proportion of the relevant content after traversing far fewer pages than comparison methods. GBS outperformed comparison methods on various categories of Web pages in the test beds, including collection of blogs, Web forums, and social networking Web site content. Further analysis revealed that both the sentiment classification module and graph-based tunneling mechanism played an integral role in the overall effectiveness of the GBS crawler.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*; I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search—*Graph and tree search strategies; Heuristic methods*

General Terms: Algorithms, Experimentation, Design, Performance

Additional Key Words and Phrases: Web crawlers, focused crawlers, sentiment analysis, opinion mining, classification, graph similarities, random walk path

ACM Reference Format:
Fu, T., Abbasi, A., Zeng, D., and Chen, H., 2012. Sentimental spidering: Leveraging opinion information in focused crawlers. ACM Trans. Inf. Syst. 30, 4, Article 24 (November 2012), 30 pages.
DOI = 10.1145/2382438.2382443 http://doi.acm.org/10.1145/2382438.2382443

A preliminary version of this article appeared in the 20th Annual Workshop on Information Technologies and Systems (WITS'10).
This research is partially funded through the National Natural Science Foundation of China (60921061, 70890084, 71025001, 91024030, and 90924302), Chinese Academy of Sciences (2F07C01), Chinese Ministry of Health (2012ZX10004801), DOD Defense Threat Reduction Agency (HDTRA-09-0058), and NSF (CNS-0709338, CBET-0730908, IIS-1236970).
Authors' addresses: T. Fu, Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043; email: tjfu@google.com; A. Abbasi, Information Technology Area, University of Virginia, Charlottesville, VA 22904; email: abbasi@comm.virginia.edu; D. Zeng (corresponding author), State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; email: zeng@email.arizona.edu; H. Chen, Department of Management Information Systems, University of Arizona, Tucson, AZ 85721; email: hchen@eller.arizona.edu.

DOI 10.1145/2382438.2382443 http://doi.acm.org/10.1145/2382438.2382443

1. INTRODUCTION

Content expressing user opinions has been proliferating in a wide array of Web 2.0 and social media applications, including online reviews, recommendations, blog articles, discussion forums, and several types of social networking Web sites. This proliferation of Web 2.0 content presents opportunities and challenges for decision-making in various domains: many political, business intelligence (BI) and marketing intelligence (MI) applications could significantly benefit from fast or even "real-time" analysis of relevant Web data. Programs called *focused* crawlers aim to create precise, task-driven Web data collections which are rich in content that meets specific requirements. The development of effective and advanced focused crawlers remains critical due to the continual need for high-quality, relevant data collections that are manageable and efficient in terms of their creation, maintenance, update mechanism, and analyzes [Pant and Srinivasan 2009].

Previous work on focused crawling has primarily emphasized the collection of topic-relevant content and ignored the embedded opinion information. However, Web 2.0 content is rich in opinion information and has obvious *sentiment polarity* (i.e., negative/positive/neutral sentiment) toward specific topics [Chen 2009; Wiebe 1994]. It has stirred much excitement and created abundant opportunities for understanding the opinions of various stakeholder and special interest groups toward social events, political movements, company strategies, marketing campaigns, and products [Chen and Zimbra 2010; Lu et al. 2010]. Companies are increasingly interested in how they are perceived by environmental and animal rights groups in terms of corporate social responsibility (CSR) [Bhattacharya et al. 2009]. Brand monitoring agencies have long sought ways to quickly "take the pulse" of consumers in regard to certain products. Knowledge of negative product sentiments can save firms millions of dollars, and in some cases, can even save human lives [Subrahmanian 2009]. For instance, pharmaceutical companies are interested in post-marketing drug surveillance (PMDS) to learn about consumer accounts of adverse drug reactions in a timely manner due to the severe legal and monetary implications [Van Grootheest et al. 2003]. As two examples of firms interested in a specific sentiment polarity surrounding a particular target, consider the cases of Nike and Kraft Foods. Nike chose to continue their relations with Tiger Woods after analysis revealed that the strong negative sentiments surrounding the endorsement deal were outweighed by the increased revenues attributable to their sponsorship of Woods [Chung et al. 2011]. Similarly, Kraft Foods engaged consultants from IBM to analyze the Web to help identify what people liked about their Vegemite product [Spangler et al. 2008].

The increasing importance of sentiment information necessitates quick and efficient focused crawler methods to collect not only topic-relevant but also sentiment-relevant content from various Web 2.0 media such as the blogosphere, social network services (SNS), video-sharing sites, forums, etc. [Liu et al. 2010]. However, there has been limited work on focused crawlers capable of effectively collecting such content. One alternative is to collect everything related to a particular topic using a traditional topical crawler, and then to identify the subset of the collection with appropriate sentiment information in an offline manner. However, this approach greatly diminishes the main advantages of focused crawlers; efficient use of computing time and storage space. Moreover, the increasing volume of online content and the need for "real-time" analysis makes such an approach less attractive, and in many cases even infeasible [Chen and Zimbra 2010; Thelwall 2007].

Actionable "real-time" intelligence requires balancing efficiency with accuracy. Focused crawlers incorporating information access refinements that improve precision without enhanced recall can be problematic. Decisions and judgements made using low-recall data collections can be heavily biased and lead to unexpectedly bad results.

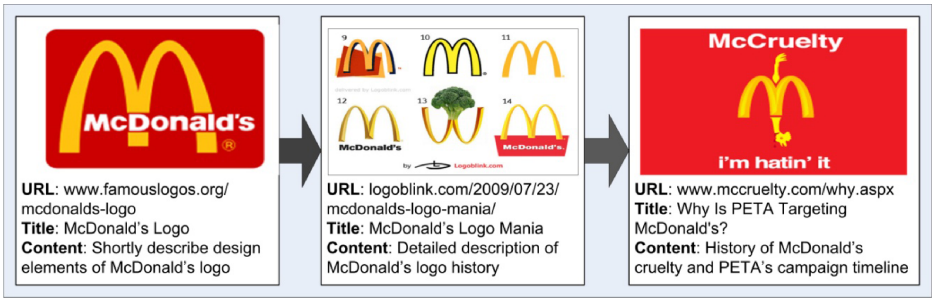

Fig. 1. Example path for tunneling.

Focused crawlers that only evaluate the out-links of relevant pages are likely to miss relevant pages that are not directly linked [Menczer et al. 2004]. This problem is exacerbated when collecting content containing specific topics and sentiments due to the lower proportion of relevant pages to irrelevant ones (as compared to traditional topic crawling tasks). Tunneling is a strategy utilized by focused crawlers to traverse irrelevant pages in order to reach relevant ones [Martin et al. 2001]. In order to attain suitable recall levels, effective tunneling is essential for focused crawlers incorporating sentiment information. Suppose McDonald's is interested in identifying and analyzing negative opinions about their brand. In Figure 1, the first page describes design elements of McDonald's logo. One of the out-links of this page provides a detailed description of McDonald's logo history. This second page also provides a link to a page from www.mccruelty.com, a Web site with strong negative sentiments towards McDonald's. This third page is obviously highly relevant to the collection task. However without tunneling, it would not be reached since a sentimental spider would only traverse the out-links of pages deemed relevant (and the first two pages are topically relevant but lack relevant sentiment).

The example presented in Figure 1 illustrates how sentiment information and tunneling can improve a focused crawler's ability to collect opinionated content. Our goal in this study is to examine whether sentiment information and tunneling is useful for crawling tasks that involve consideration of content encompassing opinions about a particular topic. We propose a novel focused crawler that incorporates topic and sentiment information as well as a graph-based tunneling mechanism for enhanced collection of opinion-rich web content regarding a particular topic. The crawler classifies Web pages based on their topical and sentimental relevance and utilizes graph similarity information in tunneling. Experimental results demonstrate the effectiveness of our crawler over several comparison focused crawlers. The remainder of the article is organized as follows. Section 2 presents a brief review of existing work on focused crawling as well as research gaps. The proposed graph-based sentiment (GBS) crawler is discussed in Section 3. Section 4 describes the experimental test bed as well as the seven comparison focused crawlers. This section also includes experimental results comparing GBS against the existing methods. Section 5 presents concluding remarks.

2. LITERATURE REVIEW

Focused crawlers aim to efficiently locate highly relevant target Web pages by using available contextual information to guide the navigation of links and are seen as a way to address the scalability limitations of universal search engines [Chakrabarti et al. 1999; Menczer et al. 2004]. Two main characteristics of focused crawlers are the contextual information and the techniques used for candidate URL ordering and classification.

Three types of contextual information are useful for estimating the benefit of following a URL: link context, ancestor pages, and web graphs [Liu 2011; Pant and Srinivasan 2005]. Link context refers to the lexical content around the URL in the page from which the URL was extracted (i.e., the parent page), which can range from text surrounding the link (called anchor text) to the whole content of the link's parent page. Ancestor pages are the lexical content of pages that lead to the parent page of the URL. Web graphs refer to the hyperlink graphs composed of in-links and out-links between Web pages.

Link context is the most fundamental contextual information in classifier-based topical crawlers and has been utilized by most prior focused crawlers [Pant and Srinivasan 2005]. The popularity of the Vector Space Model (VSM) for text classification has also resulted in the use of VSM-based crawlers that rely exclusively on link context. They have been widely used in previous studies such as Aggarwal et al. [2001], Menczer et al. [2004], and Pant and Srinivasan [2005]. A typical VSM-based crawler represents each Web page as a vector space using the TF-IDF (term frequency and inverse document frequency) weighting schema [Salton and McGill 1986]. TF-IDF vector of a candidate Web page is usually compared with vectors of relevant and irrelevant training pages in order to determine its relevance. Previous studies have also used a more selective keyword list as the basic vocabulary for the TF-IDF schema of VSM, which we refer to as Keyword-based crawler in this article [Menczer et al. 2004]. The quality of the keyword list is critical to the performance of a Keyword-based crawler. Domain experts may select keywords based on their domain knowledge. Conversely, automated feature selection techniques may be used to learn keywords that are adept at assessing the relevance of documents [Abbasi and Chen 2008; Yang and Pedersen 1997].

Crawlers that only rely on link context are often good at evaluating links of relevant pages, which is consistent with the topical locality hypothesis that claims similar content is more likely to be linked [Davison 2000]. However, the increased volume of Web data and the complex structure of the Web greatly reduces the recall of these crawlers because they fall short in learning tunneling strategies when relevant content is just a few links behind an apparently irrelevant page [Diligenti et al. 2000]. Some researchers proposed to utilize external knowledge to broaden the search space if necessary, for example to temporarily change the crawling topic from "sailing" to "water sports" based on the hierarchical relationship between words (called "hypernymy") [Martin et al. 2001]. Such relationships can be identified using the Open Directory Project (ODP) or a lexical thesaurus such as WordNet [Martin et al. 2001].

Advanced crawling techniques have been developed to overcome the shortcomings of the above crawlers by utilizing the other two types of contextual information: ancestor page and web graph. Context Graph Model (CGM) is a good example of a crawler that incorporates ancestor pages in the crawling process [Diligenti et al. 2000]. A context graph represents how a target document can be accessed from the Web and consists of in-link pages and their ancestor pages. The CGM crawler builds Naïve Bayes classifiers for each layer of the relevant training data's context graph. These classifiers are then used to predict how far away an irrelevant page is from a relevant target page. Irrelevant pages are ranked in the queue based on their perceived proximity to relevant target pages. In head-to-head comparisons, the CGM crawler outperformed several focused crawlers that rely solely on link context information [Diligenti et al. 2000].

Among the three categories of contextual information exploited by focused crawlers, web graphs rely the least on the lexical content of a page. Pattern recognition refers to the act of determining to which category or class a given pattern belongs. Based on how patterns are represented, there are two types of pattern recognition: statistical and

structural pattern recognition [Riesen and Bunke 2010]. In statistical pattern recognition, objects or patterns are represented by feature vectors. The abovementioned methods for evaluating link context and ancestor pages utilized statistical pattern recognition techniques. In contrast, structural pattern recognition utilizes symbolic data structures such as strings, trees, or graphs. Compared with feature vectors, graphs are better suited to describe spatial, temporal or conceptual relationships between objects. However, few search engines or focused crawlers have explored web graphs due to limitations in available graph information and computational constraints. Hopfield Net (HFN) [Chau and Chen 2003, 2007] models the web graph as a weighted, single-layer neural network. It applies a spreading activation algorithm on the model to improve web retrieval. HFN outperformed breadth-first search (BFS) and PageRank (which also uses web graph information) in the collection of medical Web pages. PageRank [Brin and Page 1998], which is commonly used as a baseline in focused crawling studies, simulates a random walk over the Web taken by a Web surfer and calculates the quality of a page proportionally to the quality of the pages that link to it. It attempts to identify hub nodes (i.e., pages that link many resourceful pages) in web graphs. Both HFN and PageRank use web graph information to pass accumulated weights to child pages (i.e., out-links).

Since classification techniques that using different types of contextual information have their own strengths and weaknesses, some researchers adopted ensemble techniques [Allwein et al. 2001; Schapire and Singer 1999] and implemented various voting schemes that incorporated predictions from several classifiers. For example, Fürnkranz [2002] suggested four voting schemes: majority vote, weighted sum, weighted normalized sum, and maximum confidence. However, since ensemble techniques are time-consuming, they have mostly been used to evaluate crawler results. For instance, Pant and Srinivasan [2005] used a classifier ensemble composed of eight Naïve Bayes, Support Vector Machines (SVM), and Neural Network classifiers to evaluate their crawlers.

Based on our review of prior work on focused crawling, we have identified several research gaps. To the best of our knowledge, sentiment information has never been utilized by previous crawlers. Given the proliferation of user-generated content rich in opinions and sentiments, there remains a need to evaluate the efficacy of using sentiment information for enhanced focused crawling of opinion-rich Web content regarding a particular topic. Moreover, several previous studies have pointed out that web graphs may provide essential cues about the merit of following a particular URL, resulting in improved tunneling capabilities [Broder et al. 2000; Pant and Srinivasan 2005]. However, most studies have relied primarily on link context information to inform the navigation of the focused crawler. The few studies that used web graphs relied primarily on from-to linkage relations between parent-child nodes [Chau and Chen 2003, 2007]. Web graph structure has seen limited usage. In the following section, we describe the proposed graph-based sentiment crawler (GBS) that utilizes topic and sentiment information and graph-based tunneling to identify pages containing opinions about a particular topic.

3. RESEARCH DESIGN

We propose a new focused crawler that can leverage sentiment information and labeled web graphs. This Graph-based sentiment crawler (GBS) consists of four modules: crawler, queue management, text classifier, and graph comparison. The first two modules are common to most focused crawlers. The queue management module ranks the current list of candidate URLs based on their weights. The weights associated with candidate URLs are determined by the last two modules (i.e., the text classifier and graph comparison), described in detail below. The crawler module crawls URLs in

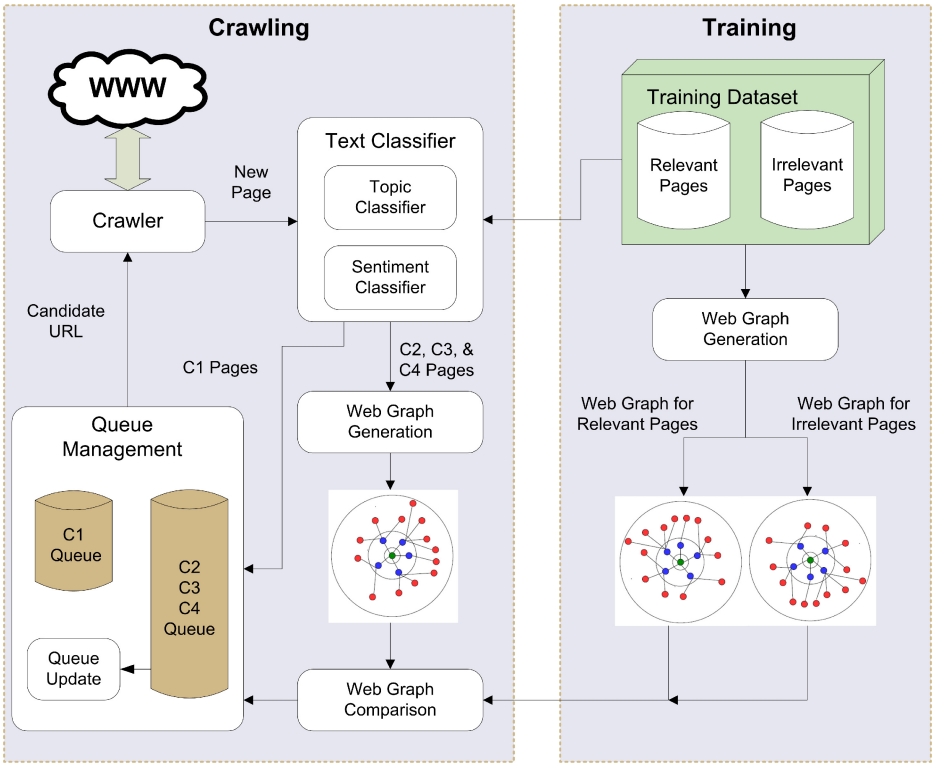

Fig. 2. System design for Graph-based sentiment crawler.

descending order based on their rank/location in the queue management module. The GBS design is shown in Figure 2.

3.1. Text Classifier Module

Our text classifier module consists of a topic classifier and a sentiment classifier. Each classifier adopts a simple, computationally efficient categorization approach suitable for use within a focused crawler. The topic classifier computes the topical relevance of a page using a trained classification model, as follows. Given a set of training pages containing known relevant and irrelevant pages, we extract all non–stop word unigrams occurring at least 3 times. Each of these keywords a is weighted using the information gain heuristic, where a weight $w(a)$ is computed based on the keyword's level of entropy reduction [Shannon 1948]. Hence,

$$w(a) = E(y) - E(y|a),$$

$$E(y) = -\sum_{i \in y} p(y = i) \log_2 p(y = i),$$

where E(y) is the entropy across the set of classes y (i.e., relevant and irrelevant pages), and

$$H(y|a) = -\sum_{j \in a} p(a = j) \sum_{i \in y} p(y = i|a = j) \log_2 p(y = i|a = j)$$

is the entropy of y given a, where $p(a = j)$ is the probability that keyword a has a value j, where $j \in \{0, 1\}$ depending on whether or not a occurs in a particular Web page. It is important to note that $E(y)$ =1 if the number of relevant and irrelevant pages in the training set are equal/balanced. For each keyword a, we also compute its relevance $r(a) \in \{-1, 1\}$, where $r(a) = 1$ if a occurs in a greater number of relevant training pages than irrelevant ones, $r(a) = -1$ otherwise. Once the topic classifier has been trained, it can be used to score a candidate page u as follows:

$$TS(u) = \sum_{a} w(a)r(a)t(a),$$

where $t(a) = 1$ if keyword a occurs in page u, $t(a) = 0$ otherwise. A candidate page u is considered topically relevant if $TS(u) > 0$.

The sentiment classifier computes a sentiment score, SS(u), for each candidate page u. The sentiment classifier considers both sentiment polarities and intensities. Sentiment polarity pertains to whether a text has a positive, negative, or neutral semantic orientation [Abbasi et al. 2008]. A given sentiment polarity (e.g., positive/negative) can also have varying intensities: for instance weak, mild, or strong [Abbasi et al. 2011]. We utilize SentiWordNet [Esuli and Sebastiani 2006], a lexical resource, to derive the sentiment polarities and intensities associated with the text surrounding relevant keywords contained in u. SentiWordNet contains three sentiment polarity scores (i.e., positivity, negativity, objectivity) for synsets composed of word-sense pairs [Esuli and Sebastiani 2006]. SentiWordNet contains scores for over 150,000 words, with scores being on a 0–1 scale. For instance, the synset consisting of the verb form of the word "short" and the word "short-change" has a positive score of 0 and a negative score of 0.75. As a preprocessing step, for each word w in SentiWordNet, we compute its semantic weight $s(w)$ as the average of the sum of its positive and negative scores across word-sense pairs. To compute $SS(u)$, we only consider the semantic weight of sentences containing relevant keywords found in u. In other words, let B represent the subset of keywords found in u where $r(b) = 1$ for each $b \in B$. Further, let K_b denote the set of words from each sentence in u that contains b. The sentiment score for each candidate page u is computed as the difference in semantic orientation between that page and the relevant pages in the training dataset, regarding the words in B. More specifically,

$$SS(u) = \frac{1}{|B|} \sum_{b \in B} \left| \left(\frac{1}{|K_b|} \sum_{i \in K_b} s(i) \right) - \left(\frac{1}{|R_b|} \sum_{j \in R_b} s(j) \right) \right|,$$

where $s(i)$ is the semantic score for word i and R_b denotes the set of words from each sentence in the relevant training pages that contains b. Candidate page u is considered to contain relevant sentiment if $SS(u)$ is less than a threshold parameter t (i.e., if $SS(u) < t$).

Figure 3 presents an illustration of the text classifier utilized by GBS. The top half of the figure shows the topic classifier, while the bottom half depicts the sentiment classifier. In the topic classifier, all keywords are indexed, weighted, and associated with one of the two classes (based on their occurrence distribution across classes). In Figure 3, keywords associated with relevant pages are denoted by circles while ones associated with irrelevant pages are depicted by squares. Each candidate page is classified as relevant/irrelevant based on the sum of the weighted presence of these keywords. The sentiment classifier computes the difference in sentiment composition between candidate page sentences containing keywords associated with relevant pages (i.e., depicted by circles) and relevant training Web page sentences containing those same keywords. Candidate pages that differ from the relevant pages by less than t are considered to

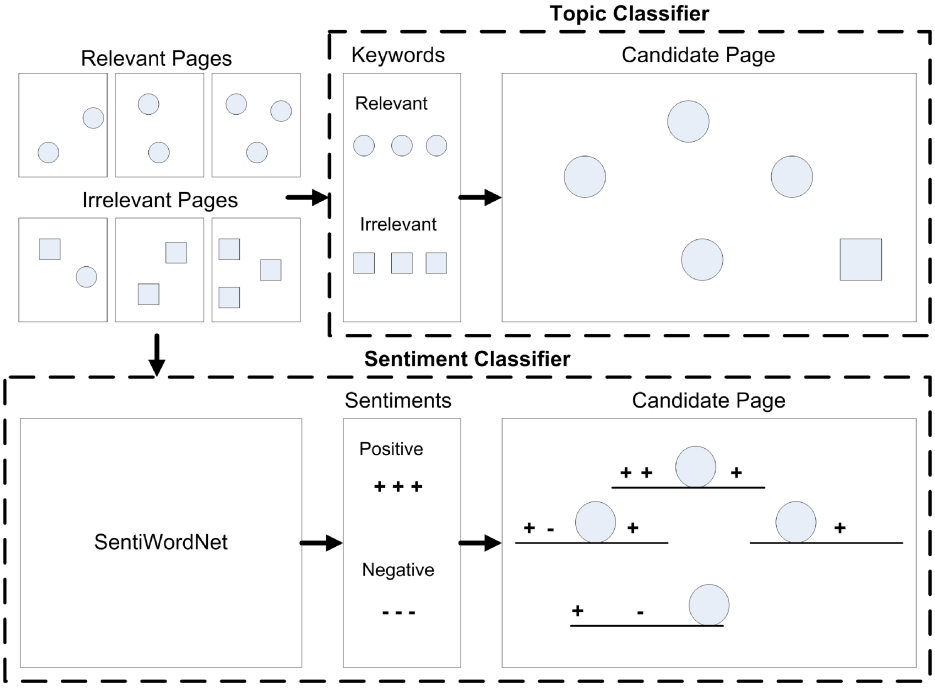

Fig. 3. Illustration of text classifier used by GBS crawler.

contain relevant sentiment information. By applying the text classifier module, each collected Web page is categorized as belonging to one of the following four classes:

C1: Relevant topic and sentiment,
C2: Relevant topic only,
C3: Relevant sentiment only,
C4: Irrelevant topic and sentiment.

Only C1 pages are considered targets of our crawler system. Previous studies have already shown the benefits of exploring links originating from targeted Web pages (i.e., out-links) [Aggarwal et al. 2001; Chau and Chen 2003, 2007; Diligenti et al. 2000]. Accordingly, the queue management module in GBS assigns C1 pages' out-links the highest weights. C2 pages are topically relevant but have irrelevant sentiment. For instance, if a company is interested in the amount of negativity surrounding a recent event, news articles describing the event (in an objective manner) would be considered C2 pages. C3 pages contain relevant sentiments but are not topically relevant. For instance, weblog and microblog pages often contain entries pertaining to an array of topics, which can diminish such pages' overall relevance to any one topic [Thelwall 2007]. Using our company-event example, a blogger may express negative sentiments regarding the event in passing (e.g., with a single entry). C4 pages are those that are not considered relevant in terms of topic or sentiment. The weights for out-links of C2, C3, and C4 pages are calculated by the graph comparison module using their labeled web graphs, described in the following section.

3.2. Graph Comparison Module

Graph matching is the process of evaluating the structural similarity or dissimilarity of two graphs and a key task of structural pattern recognition. Two broad categories

are exact graph matching, which requires a strict correspondence between two graphs or at least their subgraphs, and inexact graph matching, where a matching can occur even if there are some structural differences [Conte et al. 2004]. Inexact graph matching has received additional attention in recent years since for many applications, exact matching is impossible or computationally infeasible [Conte et al. 2004; Garey and Johnson 1979]. One of the key characteristics of inexact graph matching is the similarity measure employed. Graph edit distance, which defines the matching cost based on costs of a set of graph edit operations (e.g., node insertion, node deletion, edge substitution, etc.), is considered one of the most flexible methods and has been applied to various types of graphs [Baeza-Yates 2000; Eshera and Fu 1984; Myers et al. 2000]. However, existing methods for computing graph edit distance lack some of the formality and rigor associated with the computation of string edit distance. To convert graphs to string sequences so that string matching techniques can be used, Robles-Kelly and Hancock [2005] used a graph spectral seriation method to convert the adjacency matrix into a string or sequence order. For labeled graphs, random walk paths have been used to represent graphs as string sequences of node classes with associated occurrence probabilities [Kashima et al. 2003; Li et al. 2009]. Accordingly, the graph comparison module utilized by GBS uses random walk paths to represent the web graphs associated with candidate pages as well as known relevant and irrelevant pages. Details regarding the graph comparison module are presented in the remainder of the section.

The graph comparison module analyzes the labeled web graphs associated with pages deemed nonrelevant by the text classifier during the crawling phase to determine if they are likely to lead to relevant pages. In other words, the objective of the graph comparison module is to determine whether "tunneling" through this particular nonrelevant page could be fruitful. Algorithmically, the graph comparison module calculates the weights of C2, C3, and C4 pages in the crawler's queue based on the similarity of their discovered web graphs with those of training data, as illustrated in Figure 2.

The intuition behind the use of a graph-based tunneling mechanism is inspired by the observation that web graphs of irrelevant pages that lead to relevant content are *subgraphs* of relevant pages' web graphs. Suppose the following path leads to a targeted C1 page: C1→C2→C3→C1 (target), where the labels represent the classes associated with pages along the path. A focused crawler would explore all out-links of the seed C1 page and collect the C2 page. If it were a traditional topic-driven focused crawler, it would advance further along the path (since C2 is topic relevant) and collect the C3 page. Because this page is neither topic relevant nor sentiment relevant, the crawler would not be interested in exploring this path any further. Consequently, it would miss the targeted C1 page. To evaluate the value of irrelevant pages such as the C3 page from our example, the crawler cannot solely rely on its lexical content. However, let's assume that the path that leads to C3 (C1→C2→C3) is also quite commonly found in the web graphs associated with C1 pages. This would suggest that analyzing the out-links of the C3 page may lead to a C1 page. Hence, analysis of the similarity between web graphs of relevant and irrelevant pages may provide an estimate/indication of how close an irrelevant page is to relevant content.

As shown in the right side of Figure 2, initially we construct the web graphs of known relevant and irrelevant pages in the training dataset. A web graph, consisting of n levels of in and out links, is constructed for each page in the training data. The in-links are gathered using public in-link services such as Yahoo's site explorer inbound links API. Due to computational limitations and efficiency issues, restrictions are imposed on the number of levels employed in the web graph, as well as the number of Web pages (i.e., nodes) utilized per level. We set the level limit as 3 and sample 100 in-links for each node in the web graphs of the training data.

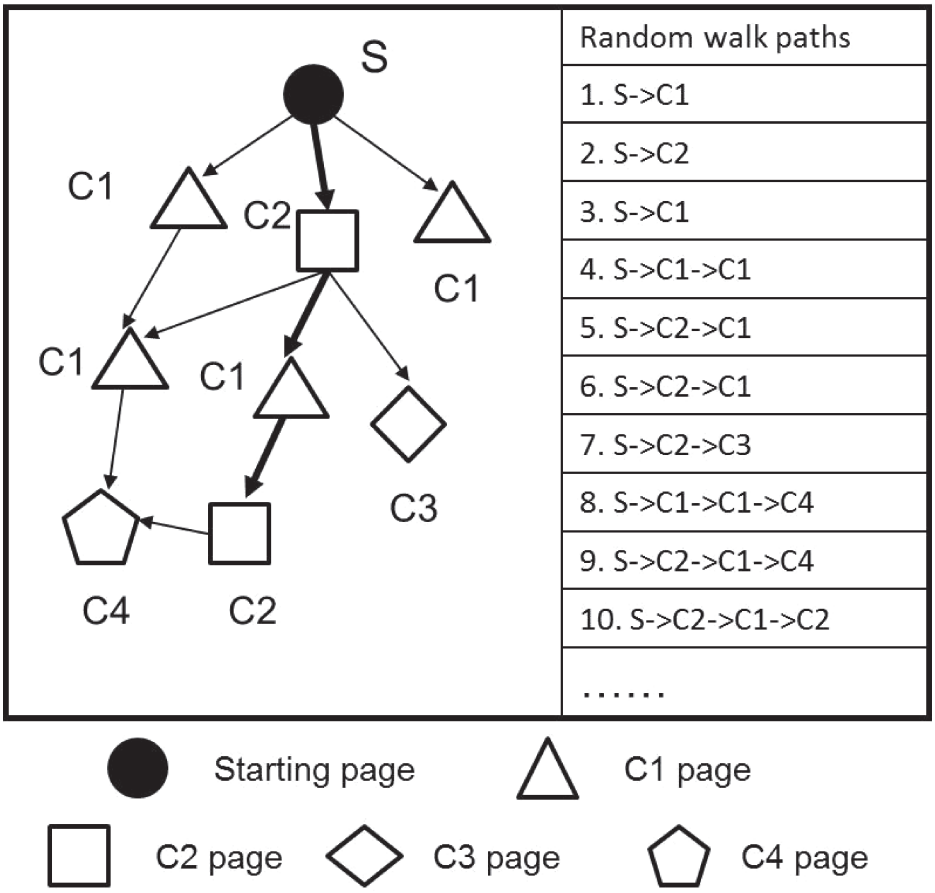

Fig. 4. Random walk paths on a labeled web graph of page S.

Nodes in the web graphs are labeled with their corresponding classes (C1-C4) using our text classifier module. Each web graph is then represented by various *random walk path* (RWP) sequences, where each RWP is composed of a series of labeled nodes that signify the traversal of a particular path along the graph [Kashima et al. 2003]. RWP sequences have been widely used in graph comparison tasks, for example patent classification using patent citation networks [Li et al. 2009]. At each step, a random walk either jumps to one of the in-links or stops based on a probability distribution. Figure 4 represents a labeled web graph of page S. Nodes in the web graph are all ancestor pages of S and their class information is depicted by various node shapes (e.g., square, triangle, diamond, pentagon). Suppose we generate RWP sequences using a 0.1 stop/termination probability and equal "jump" probabilities, the highlighted RWP sequence S→C2→C1→C2 in the middle of the graph would have an occurrence probability of 0.3 * 0.3 * 0.9 * 0.1 = 0.0081.

As illustrated in the left side of Figure 2, during the crawl, the graph comparison module is used to evaluate each C2, C3, and C4 page. The web graphs of irrelevant pages that our crawler finds in the crawling stage are constructed based on pages that have been already collected. While the maximum level of these web graphs may exceed 3, only nodes within the top 3 levels are used in the graph comparison module.

A maximum of 3 levels were used since preliminary analysis revealed that additional in-link levels increased computation times without providing meaningful performance gains. GBS generates RWP sequences for the current set of collected irrelevant pages by following their in-links. Next, the web graphs of these candidate irrelevant pages are compared against those of pages in the training data.

The similarity between two graphs is measured by the aggregated value of similarities among their RWPs multiplied by these RWPs' occurrence possibilities, and calculated using the following formula:

$$Sim(G, G') = \Sigma_h \Sigma_{h'} SimRwp(h, h') P(h|G) P(h'|G'), \quad (1)$$

where G and G' are two graphs, h and h' are RWPs of the two graphs, *SimRwp()* is used to calculate the similarity between RWPs, and *P()* returns the probability of each RWP in its graph.

As previously stated, the web graphs are composed of the four types of nodes described in Section 3.1. Moreover, the RWP sequences used are also limited to 3 hops. Therefore, the types of RWP our graph module needs to deal with are predetermined: all possible permutations of C1–C4 nodes of length 3 or less (e.g., 211, 134, 231, 31). Hence, if we use "*t*" to represent one type of RWP, formula (1) can be transformed to:

$$Sim(G, G') = \Sigma_t \Sigma_{t'} SimRwp(t, t') P(t|G) P(t'|G'),$$

where $P(t/G) = \sum_h P(h/G) Belong(h, t)$, and $Belong(h, t) = 1$ if h belongs to type t, 0 otherwise. If a type of RWP doesn't appear in one graph, *P(t|G)* returns 0.

Since web graphs of candidate irrelevant pages are assumed as subgraphs of those of relevant pages, the two RWPs used in *SimRwp()* should be of different length. In other words, RWPs originating from the candidate irrelevant pages need to be shorter than those found in our training data. In order to compare such RWPs, we employ Levenshtein distance since it is well-suited for comparisons involving data of unequal size [Levenshtein 1966]. Levenshtein distance is a metric for measuring the amount of difference between two sequences (i.e., edit distance). The Levenshtein distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character [Levenshtein 1966]. Therefore *SimRwp()* is replaced by *LD()* to represent the calculation of RWP similarity using Levenshtein distance.

If we use *SetG'* to represent the set of web graphs G' of our training data (either relevant set or irrelevant set), the web graph similarity of a candidate page with web graphs of a dataset can be calculated as an average similarity using the following formula:

$$\begin{aligned} & Sim\ (G, SetG') \\ &= Avg(\Sigma_{G'} Sim(G, G')) \\ &= Avg(\Sigma_{G'} \Sigma_t \Sigma_{t'} LD(t, t') P(t|G) P(t'|G') Short(t, t')) \\ &= \Sigma_t P(t|G) Avg(\Sigma_{G'} \Sigma_{t'} LD(t, t') P(t'|G') Short(t, t')) \\ &= \Sigma_t P(t|G) * Training(t,\ SetG'), \end{aligned}$$

where G is the web graph of a candidate page, G' is that of a training page, *SetG'* is the set of web graphs from the training data, *LD()* calculates the Levenshtein distance of two RWPs, and *Short(a, b)* returns 1 if path a is shorter than path b and 0 otherwise. *Training()* represents all the calculations that are independent of G. These calculations return the possibility for a type of path "*t*" to appear in the web graphs of the training dataset and can be done in the training stage of our graph module. During the crawling stage, our crawler only needs to calculate**,** the possibility of a type of

RWP in the discovered web graph of every candidate page. Such calculations are very fast considering the limited size of the web graphs and therefore the time complexity is definitely acceptable for crawlers.

The weight for out-links of a candidate page "m" is defined as the ratio of the page's web graph similarity score for the relevant training dataset to that for the irrelevant training dataset:

$$Weight(m) = \frac{Sim(Gm, SetG'Relevant)}{Sim(Gm, SetG'Irrelevant)}.$$

It is important to note that the web graph of a candidate URL can be updated during the crawling process when new ancestor pages (i.e., in-link pages) are discovered. Therefore the weights of candidate URLs should also be updated from time to time. In order to perform such updates in a computationally efficient manner, we update the weights of C2-C4 pages in the queue every time a predefined number of new irrelevant pages have been collected.

To the best of our knowledge, web graph similarity has not been explored in prior focused crawlers. There are two possible reasons: lack of information in the web graph structure and the time complexity issue. However, incorporating sentiment information into focused crawlers greatly enriches web graphs by providing an additional information dimension. The presence of additional node classes in the web graphs creates new opportunities for graph-based tunneling. Moreover, the time complexity for the graph comparison module utilized by GBS is computationally feasible due to the use of RWP-based inexact matching and training data that enables the use of a narrower set of promising web graph properties. In fact, recent machine learning studies have provided advanced methods to reduce the time complexity of string, tree, and graph-based matching to linear time [Rieck et al. 2010].

4. EVALUATION

In order to examine the effectiveness of the proposed GBS crawler, which utilizes sentiment information and a labeled web graph, experiments were conducted that compared the system against traditional topic-driven crawlers, including Vector Space Model (VSM), Keyword-based method, Context Graph Model (CGM), Hopfield Net (HFN), PageRank and Breadth-First-Search (BFS) [Aggarwal et al. 2001; Brin and Page 1998; Chau and Chen 2003, 2007; Diligenti et al. 2000]. BFS was included since it is often used as a benchmark technique in focused crawling studies [Chau and Chen 2007; Pant and Srinivasan 2005]. The other techniques incorporated are representative of those that adopt the aforementioned three types of contextual information: link context, ancestor pages, and web graph information [Pant and Srinivasan 2005].

4.1. Test Bed

Because of the dynamic nature of the Web [Arasu et al. 2001; Cho and Garcia-Molina 2003], we created a controlled environment for our experiments by taking a snapshot of a portion of the Web. A controlled environment was necessary to ensure that GBS and various comparison crawlers, each of which were run several times under varying experiment conditions, were evaluated on the exact same test pages [Menczer et al. 2004; Pant and Srinivasan 2005]. We used two test beds for evaluation: animal rights content of relevance for corporate social responsibility (CSR); and drug and medicine content of relevance for post-marketing drug surveillance (PMDS).

For the CSR test bed, we aimed to collect content containing negative sentiments towards organizations considered to infringe on animal rights and/or animal protection initiatives. Such content sheds light on an important and active constituency that

exerts considerable influence on the political and corporate landscapes [Bhattacharya et al. 2009]. The test bed also contained content with neutral or opposing sentiments, such as objective information and news about these groups, as well as criticism targeted towards animal rights activists by individuals and groups holding opposing views. The variety of content in the test bed made it suitable for our experiments.

The PMDS test bed was geared towards identifying content containing negative sentiments regarding various drugs and medicines. Pharmaceutical companies are increasingly interested in learning about negative consumer opinions towards their products since such sentiments may be important indicators of potential adverse drug reactions and/or drug-drug interactions [Van Grootheest et al. 2003]. Given the severe social, legal, and monetary implications of unsafe drugs, near real-time post-marketing surveillance of consumer drug experiences is of great importance to multiple stakeholder groups [Brewer and Colditz 1999].

Both test beds were built by collecting up to 6-levels of out-links from a set of starting URLs. For the CSR test bed, we used the homepages of 145 animal rights activist groups (e.g., Animal Liberation Front (ALF), People for the Ethical Treatment of Animals (PETA), etc.) as the starting URLs. The resulting test bed contained 524,338 Web pages with a size of about 25 GB. For the PMDS test bed, we used 100 homepages pertaining to drug-related Web sites as the starting URLs (e.g., www.drugs.com, www.rxlist.com, etc.). The collected test bed contained 12,362,406 Web pages with a size of approximately 561 GB. Both test beds included pages from Web sites, forums, blogs, and social networking sites.

For both test beds, to train the text classifier and graph comparison modules of GBS, we built a training dataset that consisted of 800 target/relevant Web pages and 800 irrelevant ones (i.e., 1600 training pages per test bed). For each test bed, these pages were manually selected from the test beds and the WWW by two independent domain experts. In other words, two individuals with expertise in animal rights and two with extensive knowledge of prescription drugs developed the training sets. Consistent with prior work [Pant and Srinivasan 2005], in addition to the GBS modules, this data was used to train two accurate but computationally expensive gold standard support vector machines (SVM) classifiers that used over 10,000 learned attributes from each of the two 1,600 Web page training sets [Abbasi and Chen 2008]. The two gold standard classifiers' attributes encompassed word n-grams, parts-of-speech tag n-grams, as well as various lexical and syntactic measures. To evaluate the gold standard classifiers, the domain experts labeled 2,000 randomly selected pages from each test bed as relevant or irrelevant. The gold standard classifiers attained 89.4% and 90.5% accuracy, respectively, on these 2,000 evaluation pages from the animal rights and prescription drug test beds. Table I shows a breakdown of the gold standard classifiers' performance results on the two 2,000 page evaluation sets. The table depicts percentage overall accuracy and class-level recall, while numbers in parentheses indicate page quantities. Both classifiers were fairly balanced in terms of their recall performance on relevant and irrelevant pages. These two classifiers were applied on the entire test beds to construct our gold standard. With an average run time of 0.5 seconds per page, including time needed to extract values for 10,000 text attributes and perform classification, the SVM classifier took over a month to process the (larger) drug test bed when run in parallel on 2 high-end servers. In contrast, the text classifier employed by GBS had an average run time of 0.02 seconds per page (25 times faster), making it more suitable for near real-time analysis. It is important to note the relationship/mapping between relevant and irrelevant pages, as classified by the gold standard classifier, and the C1-C4 classifications made by the GBS text classifier. C1 pages would be relevant while C2-C4 would be irrelevant pages with potential to enrich the graph utilized by the GBS tunneling module.

Table I. Performance of Gold Standard Classifiers on 2,000 Test Pages

Test Bed	Overall Accuracy	Relevant Recall	Irrelevant Recall
Corporate Social Responsibility (CSR)	89.4% (2000)	90.2% (510)	89.1% (1490)
Post-marketing Drug Surveillance (PMDS)	90.7% (2000)	90.4% (396)	90.8% (1604)

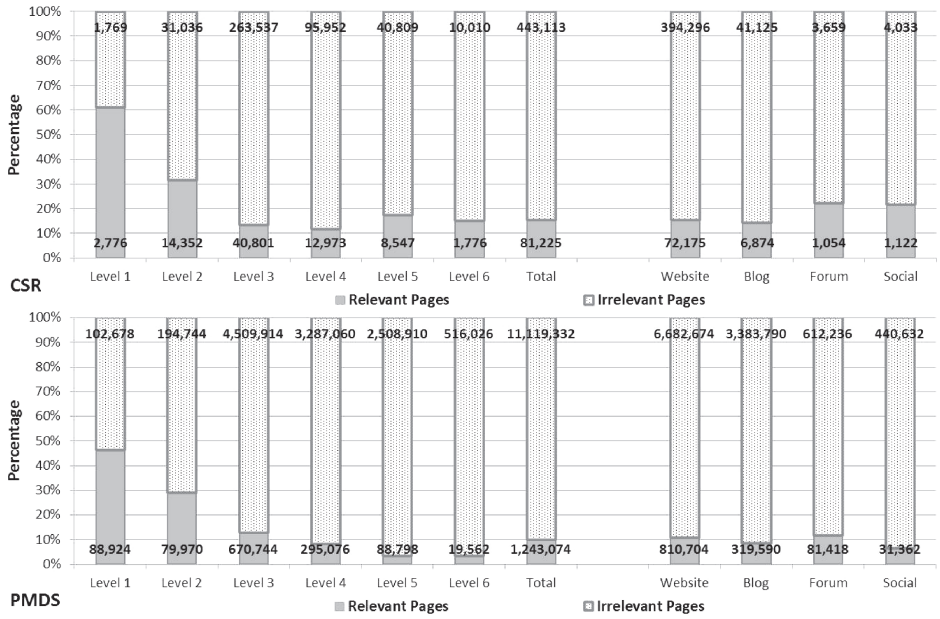

Fig. 5. CSR and PMDS test bed statistics (grouped by level and by category).

Figure 5 shows a level-by-level breakdown of the number of relevant and irrelevant Web pages in the two test beds (CSR-top, PMDS-bottom), based on the gold standard classifiers. Here, level 1 pages refer to the out-links of the starting URLs, while level 2 pages are the level 1 pages' out-links. The numbers displayed on each bar chart represent the number of relevant/irrelevant pages. For example, in the CSR test bed, at level 1, there were 2,776 relevant and 1,769 irrelevant pages (i.e., more than 60% of all the level 1 out-link pages are relevant). Not surprisingly, the percentage of relevant pages tended to decrease as we moved further away from the starting URLs. This is why relevant pages in levels that are further out from the seed URLs pose difficulties for traditional focused crawlers; their successful collection often necessitates traversal of irrelevant in-link pages. In total, only 15% of the CSR pages (81,225) and approximately 10% of the PSD pages (1.24 million) were classified as relevant.

Figure 5 also shows the number of pages in the test bed associated with four categories: blogs, Web forums, social networking sites, and standard Web sites (i.e., non-user-generated content). Interestingly, user-generated content (i.e., blogs, forums, and social networking Web sites) comprised only 11% of the CSR pages, but encompassed 40% of the PMDS test bed. Analysis of the data revealed that the CSR pages included hundreds of smaller nonprofit organizations (i.e., standard Web sites with an ".org" domain) that serve as vehicles for collective action (e.g., through community newsletters and articles expressing important sentiments). In contrast, given the more individualistic and personalized nature of drug-related sentiments, user-generated content

was more abundant. In particular, 30% of the PMDS test bed was blog pages. The number of forum and social networking Web site pages was also lower due to the fact that our collection was constrained to the surface web. Forums or social networking pages requiring login information (i.e., the deep/hidden web) could not be collected [Fu et al. 2010]. Overall, the two test beds embodied several interesting differences, including variations in size, composition across page categories, percentage of pages deemed relevant, and contrasting topics covered. These differences had important implications for the experiment results, as discussed later on in Section 4.4.

4.2. GBS Text Classifier and Graph Module Training

The GBS text classifiers depicted in Figures 2 and 3 used the 1,600 training pages to learn features and weights. A separate set of text classifiers (i.e., one topic and one sentiment classifier) were trained for each of the two test beds. The topic classifiers extracted keywords from the training documents and assigned those keywords information gain weights which were used to assess the topical relevance of candidate pages during the crawl. Similarly, the sentiment classifiers compared test bed pages against the relevant and irrelevant training pages to assess sentiment match when spidering. The combination of topic and sentiment relevance was used to determine the classification of candidate pages encountered during the collection process (e.g., C1-C4).

The GBS graph module training was performed as follows. We used a public in-link service to collect up to 3 levels of in-link pages for the 1,600 training pages. The labeled web graphs of these training pages were used to learn random walk path (RWP) sequences for our graph-based tunneling module. The in-link graph pages were labeled using the text classifier module described in Section 3.1, which assigned each page a label of C1-C4. As shown in formula (2), training of the graph module focuses on the function *Training()*, which returns the probability that a particular RWP sequence will appear in the web graphs associated with the training data. Since the in-link web graphs used by GBS did not exceed 3 levels, we only considered RWP sequences with a maximum length of 3 hops, excluding the target page. Since the purpose of tunneling is to traverse irrelevant pages with the hope of reaching relevant ones, each of the RWP sequences originated from irrelevant pages (i.e., ones labeled as C2, C3, or C4). This resulted in 60 possible RWP sequences.

Table II lists the top 20 RWP sequences based on the ratio of their probabilities of appearing in the web graphs of relevant training pages as compared to irrelevant ones on the CSR (i.e., animal rights) test bed. Each RWP sequence is represented by the labels corresponding to the graph nodes comprising that particular sequence. A given sequence XYZ can be interpreted as a candidate page X, with a level-1 inlink Y and level-2 inlink Z. For example, RWP "211" refers to a path originating from a C1 page, and subsequently going through another C1 page before reaching a candidate C2 page. The number "5" is used to denote a sequence with an early termination. Hence, RWP sequences that end with the number "5" are RWPs with a length of two. For example RWP "215" is a two-node path in which a C1 page points to a candidate C2 page. The example path shown in Figure 1 (in Section 1) can be viewed as a successful 225 that leads to target content. The relevant and irrelevant possibility columns indicate the observed likelihood, on the training data, for the given RWP path to point to relevant or irrelevant pages. In other words, the RWP "211" points to relevant (i.e., C1) pages on the next hop 37.16% of the time. Based on the results, RWP sequences that begin with "21" are pervasive at the top of the table. In other words, paths ending with a C1 pages pointing to a C2 page are very likely to point to additional C1 pages (i.e., those considered relevant). Conversely, while the RWP sequence "411" has the second

Table II. Top 20 RWPs for CSR Test Bed Based on Graph Module Training

RWP	Relevant Possibility	Irrelevant Possibility	Rel Pos /Irr Pos	RWP	Relevant Possibility	Irrelevant Possibility	Rel Pos /Irr Pos
211	0.3716	0.1828	2.0327	225	0.1384	0.1027	1.3473
212	0.2779	0.1557	1.7845	222	0.1360	0.1032	1.3169
311	0.3181	0.1788	1.7793	411	0.3195	0.2533	1.2614
215	0.2752	0.1549	1.7772	331	0.1697	0.1426	1.1899
221	0.2583	0.1517	1.7028	313	0.1736	0.1467	1.1838
213	0.2274	0.1532	1.4846	214	0.2297	0.2199	1.0449
312	0.2270	0.1537	1.4768	412	0.2284	0.2276	1.0037
321	0.2238	0.1572	1.4240	232	0.1164	0.1187	0.9805
231	0.2136	0.1546	1.3820	421	0.2253	0.2318	0.9721
315	0.2040	0.1494	1.3653	241	0.2156	0.2239	0.9628

highest relevant possibility, it is still not highly ranked due to the fact that its irrelevant possibility is also high. The last three shaded RWP sequences have a ratio less than 1, which suggests that they are more likely to link to irrelevant pages. Consequently, the graph-based tunneling module employed by GBS incorporated only those RWP sequences with a relevant-to-irrelevant ratio greater than 1 (i.e., the top 17 paths in Table I).

Of the 17 RWP sequences with a relevant-to-irrelevant ratio greater than 1, nearly two-thirds contain at least one C2 page. Analysis of the test bed reveals that two types of C2 pages are quite common in the RWP sequences. The first are news articles/reports from news Web sites such as CNN, MSNBC, and BBC. These pages usually describe stories and facts in an objective manner with neutral sentiment. The second are pages composed of sentiments that oppose the opinions/views inherent in the relevant pages. Prior research has noted that the highly interactive nature of Web 2.0 media results in linkages between content composed of diverse and often opposing opinions [Tremayne et al. 2006]. For examples, bloggers who argue with others often provide links to their opponents' articles in their own blog entries in order to justify their arguments. Similarly, special interest groups often point to content associated with organizations and groups that do not share (or in some cases even oppose) their views, beliefs, and philosophies [Fu et al. 2010; Thelwall 2007]. Equally interesting are the remaining one-third of the RWP sequences, which do not contain any C2 pages. These sequences are primarily anchored by C3 pages: ones that are not topically relevant but that have relevant sentiment. These are Web pages that do not discuss the topic of interest in detail, but mention a few keywords in passing, with the appropriate sentiment polarity/intensity. Such sequences are important since a traditional topic crawler would have difficulty traversing C3 pages.

Table III lists the top 20 RWP sequences on the PMDS (post-marketing drug surveillance) test bed. Interestingly, 14 of the sequences are the same as those ranked in the top 20 on the CSR test bed. Once again, RWP sequences with C2 and C3 candidate pages are prevalent, suggesting that certain in-link paths may be useful for tunneling across different test beds, irrespective of the application task/domain. In the context of PMDS, common C2 pages include ones from medical Web sites such as WebMD as well as positive online reviews. Whereas 8 of the top 17 RWP sequences for CSR contain at least one C3 page, 13 of the top 19 RWPs for PMDS have at least one C3 page. This suggests that pages which contain relevant sentiment, but are not deemed topically relevant, may play an even bigger role in the graph-based tunneling module in the context of the PMDS test bed.

Table III. Top 20 RWPs for PMDS Test Bed Based on Graph Module Training

RWP	Relevant Possibility	Irrelevant Possibility	Rel Pos /Irr Pos	RWP	Relevant Possibility	Irrelevant Possibility	Rel Pos /Irr Pos
212	0.3819	0.2001	1.9082	233	0.3039	0.2355	1.2907
211	0.2445	0.1286	1.9017	213	0.1989	0.1547	1.2861
312	0.1763	0.1096	1.6093	222	0.2987	0.2429	1.2298
231	0.1415	0.0965	1.4658	232	0.3444	0.2806	1.2272
321	0.1363	0.0961	1.4185	311	0.1428	0.1205	1.1854
332	0.2742	0.1956	1.4023	221	0.1058	0.0917	1.1547
313	0.2586	0.1851	1.3972	322	0.2903	0.2522	1.1512
315	0.2172	0.1592	1.3638	225	0.3163	0.2826	1.1192
331	0.1957	0.1491	1.3119	215	0.3520	0.3378	1.0421
223	0.1717	0.1315	1.3056	413	0.1906	0.1952	0.9765

4.3. Experiment Setup

Seed URLs have important implications for the performance of focused crawlers. Consequently prior studies have advocated the use of meaningful, relevant seed URLs [Menczer et al. 2004; Srinivasan et al. 2005]. Accordingly, for each test bed, we identified a set of 500 relevant URLs from the test collection. These URLs were intentionally different from the starting URLs used to collect the test beds, and also different from the ones used for training (i.e., the URLs domains for the 500 URLs in the seed pool were entirely different). In the experiments, 10 bootstrap runs were conducted where 200 of the 500 relevant URLs were randomly selected in each run, and used as the seed URLs for that run (by each of the crawlers). Hence, GBS and all comparison algorithms were run 10 times, using 200 seed URLs. The averaged results across the 10 runs were used to evaluate performance.

All comparison techniques were run using the best parameter settings, which were determined by tuning these methods' parameters on the actual test bed data. Most of the parameter values used were consistent with prior research. The VSM and Keyword methods rely on link context for URL navigation [Menczer et al. 2004; Pant and Srinivasan 2005]. The two TF-IDF vectors for VSM contained all the words that appeared more than twice in the relevant and irrelevant training pages [Aggarwal et al. 2001]. In contrast, the two vectors for the Keyword method contained fewer words since these words were selected using the information gain heuristic. For both VSM and Keyword, candidate URLs were weighted based on the ratio of cosine similarities between their vectors and the vectors of relevant and irrelevant training data. For CGM, four Naive Bayes classifiers were constructed for targeted content and 3-level in-links (i.e., the context graph), respectively, using the training corpus [Diligenti et al. 2000]. Every Web page retrieved by CGM was represented by a reduced TF-IDF vector (relative to an identified vocabulary based on the training corpus), and was classified/assigned to a queue corresponding to the most probable layer of the context graph (i.e., target or levels 1–3) based on the four classifiers' predictions. For HFN, we followed the implementation employed by Chau and Chen [2003, 2007] by using two sets of key phrases selected using the information gain heuristic from Web page content and anchor texts. The parameter values used in HFN's spreading activation algorithm were similar to those suggested by Chau and Chen [2007]. A Naïve Bayes classifier-based crawler (NB) was also utilized as a comparison method [Diligenti et al. 2000; Pant and Srinivasan 2005]. The NB classifier was constructed using the 1,600 training pages. As with other comparison methods, NB was run using settings that resulted in the best performance on the test data. The NB feature set was composed of

the top word n-grams (unigrams, bigrams, and trigrams) learned from the training set using the information gain heuristic. This feature set was chosen since it was found to provide better performance than simply using bag-of-words or learned unigrams. PageRank, which has commonly been used as a baseline method, was applied using the approach described by Cho et al. [1998] and Chau and Chen [2003]. The crawler was run using a best-first search with PageRank used as the ranking heuristic. In other words, the spider firstly retrieved pages in the queue with the highest PageRank. Consistent with previous studies, we set its damping factor to 0.9 [Chau and Chen 2007; Menczer et al. 2004]. Prior work has often computed PageRank after certain intervals during the crawl, as new information regarding the web graph (i.e., additional nodes and links) becomes available [Menczer et al. 2004]. While this setup is more realistic since complete PageRank information may not be available in advance, in order to present the best performance results, we pre-computed the PageRank for each page in the test bed on the entire collection and used these pre-computed values during the PageRank crawl.

The evaluation metrics used to assess performance were F-measure, precision, and recall:

$$F\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall},$$

where Precision = (number of relevant Web pages retrieved)/(total number of Web pages retrieved); Recall = (number of relevant Web pages retrieved)/(total number of relevant pages available in the target set). Both recall and precision have been widely used in previous focused crawling studies [Menczer et al. 2004; Pant and Srinivasan 2005].

In the following section, we describe the results for three experiments. In the first experiment, we evaluated the proposed GBS crawler in comparison with the seven comparison methods: VSM, Keyword, CGM, NB, HFN, BFS, and PageRank. All methods were run on the test bed described in Section 4.1. The results for this experiment are presented in Sections 4.4 and 4.5. In the second experiment, we evaluated GBS and comparison methods on six subtopics within the PMDS and CSR test beds (Section 4.6). The third experiment evaluated the importance of the sentiment classifier and labeled graph-based tunneling module utilized by GBS to the methods overall effectiveness (Section 4.7).

4.4. Experiment Results

Figure 6 shows the F-measure, precision, and recall trends for GBS and the seven comparison methods on the 528 thousand Web page CSR test bed (left column) and 12.4 million Web page PMDS test bed (right column). The y-axis depicts the average percentage F-measure/precision/recall across the 10 bootstrap runs, while the x-axis displays the number of pages collected at that point in the crawl. Only five of the methods (GBS, CGM, NB, BFS, and PageRank) traversed the entire collection. In contrast, HFN, Keyword, and VSM all used a stopping rule. With respect to the precision and recall trends, the x-axis displays the number of pages collected at that point in the crawl. The y-axis displays the precision (middle of Figure 6) and recall (bottom of Figure 6). Precision was computed as the percentage of collected pages that were relevant [Menczer et al. 2004; Pant and Srinivasan 2005]. Recall was computed as the percentage of total relevant pages collected at that point. Therefore, the recall values for all methods converged towards 100% as the total number of pages collected increased.

With respect to the CSR test bed (left column), GBS and CGM had the best overall performance. While these two techniques had similar F-measures on the first 50K pages, GBS performed considerably better than all comparison methods on the

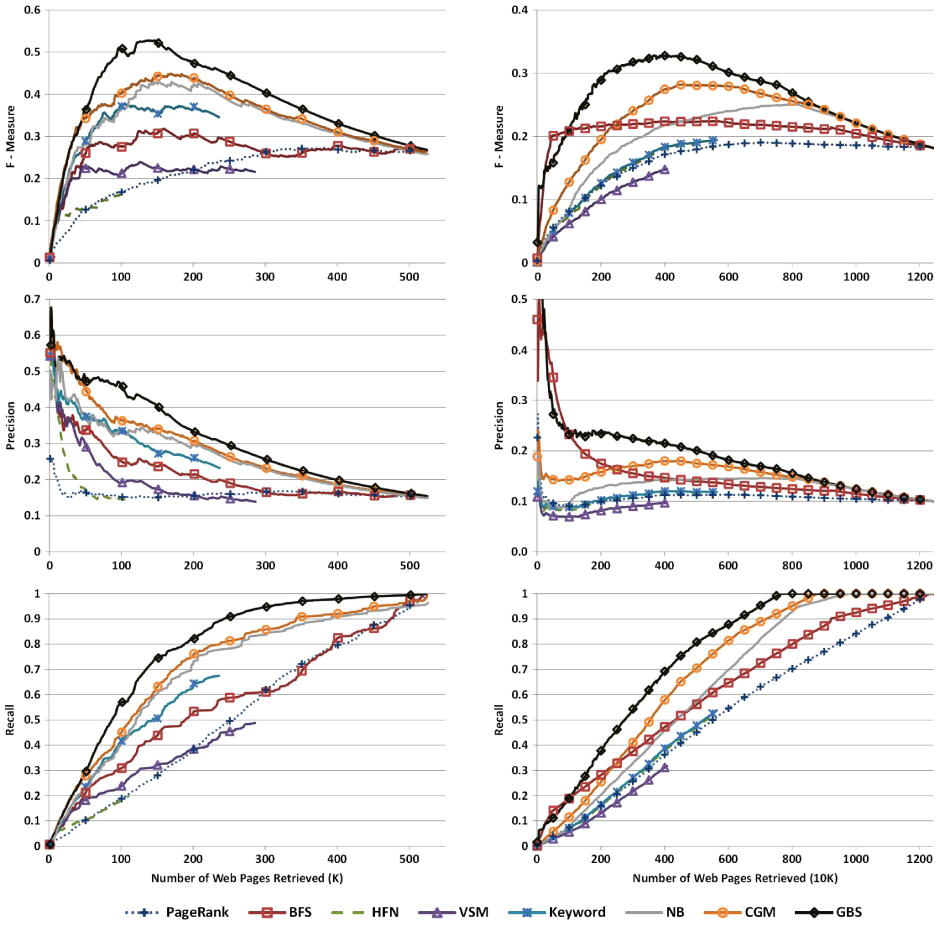

Fig. 6. F-Measure, Precision, and Recall trends for GBS and comparison methods on CSR (left) and PMDS (right) test beds.

remainder of the pages, with F-measure values exceeding 50%. With respect to the remaining comparison methods, NB, Keyword, and BFS had the best performance, followed by VSM, HFN, and PageRank. PageRank's poor performance is consistent with prior studies that have also noted that the method is less effective when applied to focused crawling tasks [Chau and Chen 2007; Menczer et al. 2004]. BFS performed well, with an average F-measure close to 0.27%. It possibly benefited from the fact that two-thirds of relevant pages in the test bed were within the network of 3-level out-links. HFN stopped crawling at a very early state (about 70k pages) since it used a stopping rule that depended on the number of relevant phrases found in retrieved Web pages' body and anchor text [Chau and Chen 2007]. The precision and recall results reveal that the enhanced performance of GBS was balanced; it outperformed all comparison methods in terms of both precision and recall. With respect to the comparison methods, the results were also consistent with CGM, Keyword, and BFS having the best precision and recall trends. For most techniques, precision decreased as the number of pages collected increased. This is not surprising since the proportion of relevant pages was greater in levels 1–2 of the test bed. Hence, as the crawlers went further out, their precision rates decreased since the number of potentially relevant pages

Table IV. Standardized Area Under the Curve (AUC) Values

Technique	CSR Test Bed			PMDS Test Bed		
	F-Measure	Precision	Recall	F-Measure	Precision	Recall
GBS	0.3915	0.3199	0.7888	0.2597	0.1914	0.7568
CGM	0.3481	0.2820	0.7193	0.2238	0.1490	0.6967
NB	0.3321	0.2636	0.6960	0.1989	0.1293	0.6448
Keyword	0.3204	0.3260	0.4186	0.1378	0.1079	0.2584
BFS	0.2673	0.2150	0.5751	0.2027	0.1487	0.6246
PageRank	0.2194	0.1604	0.5151	0.1620	0.1067	0.5323
VSM	0.2101	0.2133	0.2996	0.0943	0.0830	0.1431
HFN	0.1289	0.2131	0.1138	0.1314	0.1037	0.2397

subsided. From the early onset, GBS had recall rates that were at least 10%–15% higher than the best comparison methods (CGM and Keyword), and 25%–30% greater than the next best methods: BFS and VSM. This performance gain has important implications for real-time business and marketing intelligence. GBS was able to collect a high proportion of the relevant pages much faster than the comparison methods. Case in point, GBS collected 50% of the relevant pages after traversing only 79,000 pages. In contrast, Keyword and BFS had to traverse 144,000 and 190,000 pages (i.e., 65,000 and 111,000 more pages) respectively, in order to reach 50% recall.

On the PMDS test bed (right column), GBS once again had the best overall performance with respect to f-measure, precision, and recall. With respect to the comparison methods, CGM had the best overall performance, followed by BFS and NB. While BFS had marginally better performance on the first 10% of the collection, the GBS crawler had considerably better precision and recall for the remainder of the crawl. Consequently, GBS was able to attain 50% recall (i.e., collect half of the relevant pages) after traversing nearly 1 million fewer pages than the next best method, CGM, and approximately 1.6 million fewer pages than BFS. Content-based methods such as Keyword and VSM did not perform as well on the PMDS test bed; VSM was even outperformed by the PageRank baseline.

Table IV shows the area under the curve (AUC) results corresponding to the F-measure, precision, and recall trends presented in Figure 6. Since three of the methods did not traverse the entire collection, the AUC values were standardized to a 0-1 scale by dividing them by the total number of pages collected. Based on the table, it is evident that GBS had the best AUC values for F-measure and recall on both test beds. While Keyword performed marginally better in terms of its precision AUC value on CSR, this enhanced precision was coupled with significantly lower recall. The results presented in Table IV, along with Figure 6, suggest that GBS is well suited for focused crawling tasks involving topic and sentiment information.

We conducted level-based analysis on the CSR test bed to see how each method performed at different levels of the test bed [Diligenti et al. 2000]. Since the seed URLs used in the 10 bootstrap runs were considered level 0 pages, all out-links of the seed pages were considered level 1, while those pages' out-links were level 2 (and so on). The out-link level for each page was averaged across the 10 bootstrap runs, relative to the seed URLs for each run, and rounded to the nearest whole number. Figure 7 shows the average recall trends across bootstrap runs for pages at levels 1–6 of the CSR test bed. The results reveal that GBS performed well at all levels. It had the best recall values on levels 3, 4, 5, and 6, while BFS had better performance on level 2. The enhanced recall of GBS on pages in deeper levels was a critical factor in its overall performance. While not depicted, GBS attained similar enhanced performance on levels

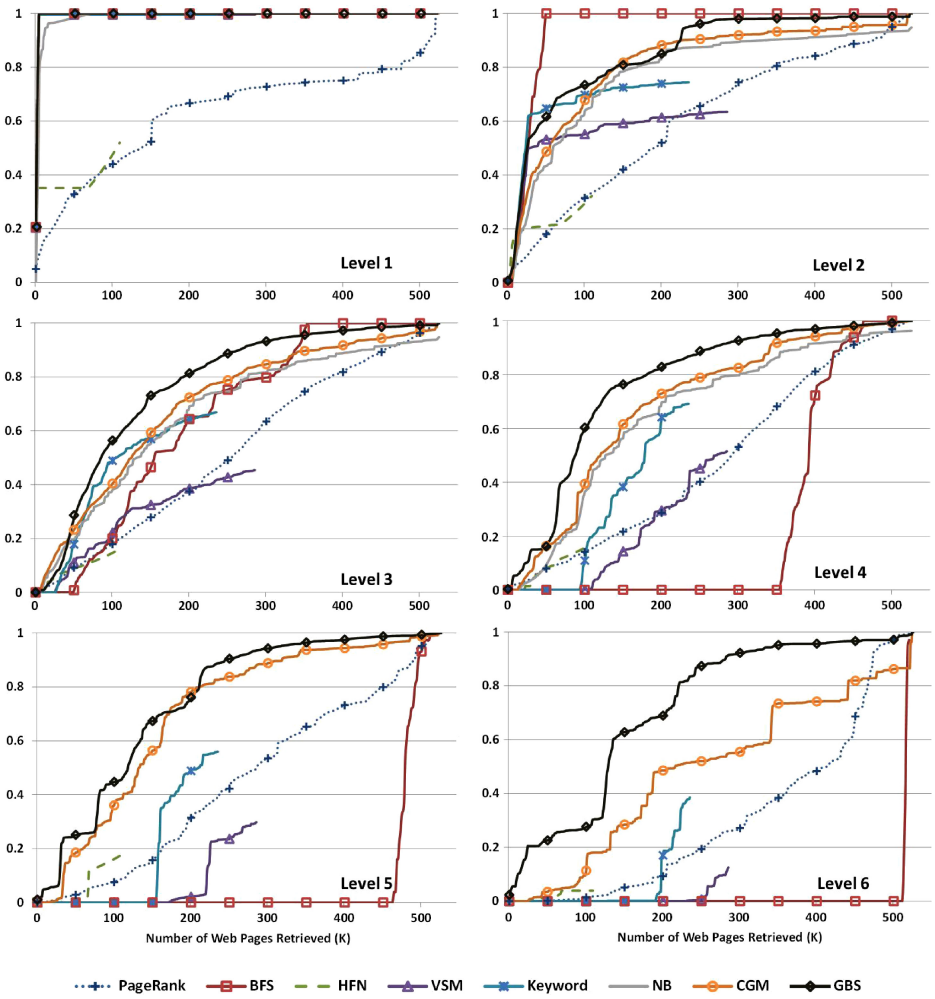

Fig. 7. Recall trends for pages at levels 1–6 of CSR test bed.

3–6 of the PMDS test bed. The results support the notion that the graph-based tunneling mechanism and sentiment classifier utilized by GBS can improve focused crawling capabilities for tasks involving the collection of opinionated content on a particular topic.

4.5. Impact of Web Page Categories

We analyzed the effectiveness of GBS and the comparison methods on the four test bed page categories: blogs, forums, social networking sites, and standard (i.e., non-user-generated) Web sites. Using the crawl data from the results presented in the previous section, recall rates were computed for each of the four aforementioned page categories. Figures 8 and 9 show the recall trends on the CSR and PMDS test beds, respectively. GBS had the best performance on all four page categories, on both test beds. With respect to the comparison methods, CGM, NB, and BFS had the best performance. With respect to the user-generated content page categories, the performance gains for GBS were most pronounced on the blog and social networking Web site pages. On the

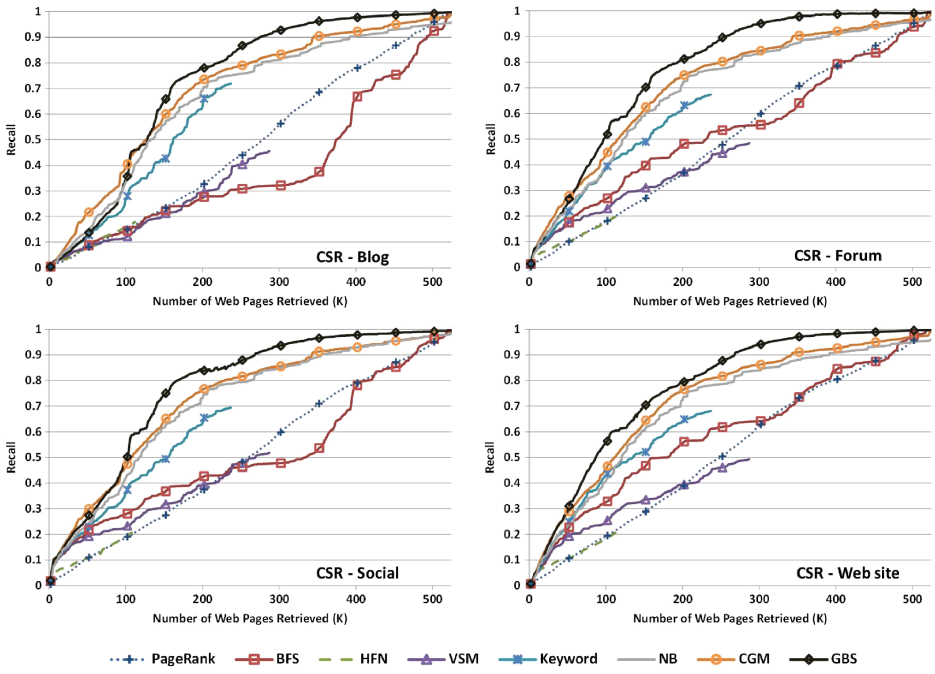

Fig. 8. Recall trends for different page categories in the CSR test bed.

PMDS test bed, BFS performed very well on the forum pages. This was due to the fact that 75% of the forum pages in the PMDS test bed were within 3 out-link levels from the seed URLs. Conversely, the majority of user-generated content pages in both test beds typically tended to be further away. For instance, more than 60% of blog pages were at out-link levels 4–6 on both test beds. Similarly, 50% or more of the social network Web site pages were further than out-link level 3. As a result, GBS and CGM had the best performance on these two categories on both CSR and PMDS. In the case of PMDS, GBS and CGM's performance was even markedly better than NB. This gain is attributable to the use of tunneling. On the larger test bed, tunneling facilitated enhanced identification of content further away from the seed URLs. Despite the fact that more than half the blog and social networking pages were three or more hops out, GBS was able to collect 75% of the relevant blog and social networking site pages in the first third of the crawl (i.e., within the first 400 million pages). With respect to the remaining comparison methods, Keyword had the best performance. HFN had decent performance on blogs and social networking Web site pages on the PMDS test bed. Consistent with the overall results, PageRank and VSM performed poorly across all four page categories.

Table V shows the AUC values for recall. GBS outperformed the best comparison method by 3%–7% on all four page categories, for both test beds. With respect to comparison methods, CGM and NB had the best performance. Interestingly, GBS, CGM, and NB all had their lowest recall rates on the forum pages. Analysis of the test bed revealed that forum pages were the most likely to contain diverse opinions and topics within a single thread/discussion page. In contrast, the blog and social networking site pages were relatively more homogenous with respect to their topic and sentiment composition. Consequently, more relevant forum pages were misclassified as C2, C3, and (to a lesser extent) C4 pages by the GBS text classifier than blog and social networking

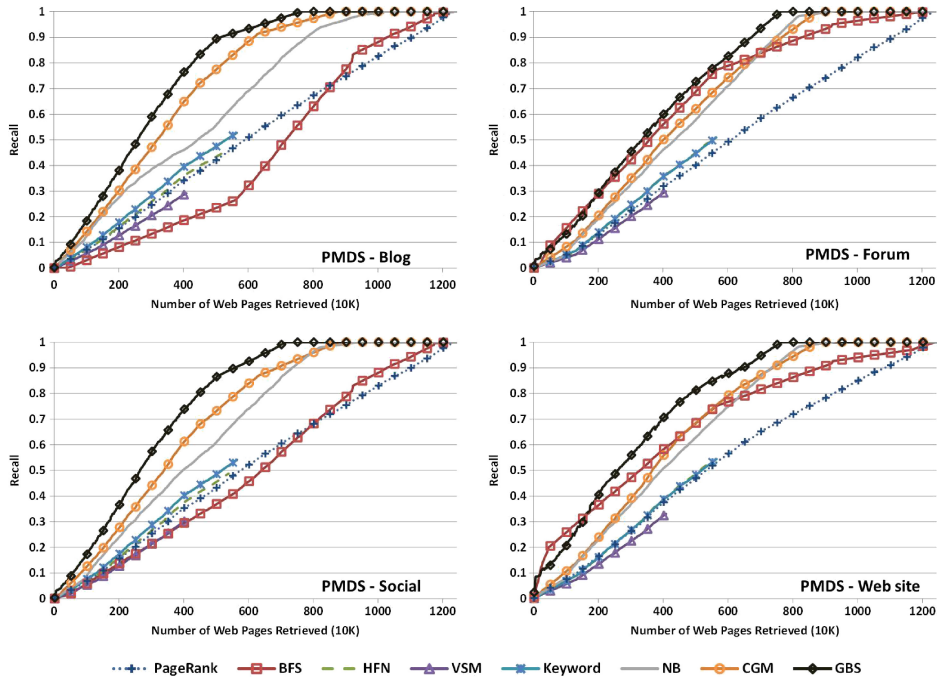

Fig. 9. Recall trends for different page categories in the PMDS test bed.

Table V. Standardized Area Under the Curve (AUC) Values for Recall on Different Page Categories

Technique	CSR Test Bed				PMDS Test Bed			
	Blog	Forum	Social	Web site	Blog	Forum	Social	Web site
GBS	0.7722	0.7263	0.7642	0.7725	0.7767	0.7168	0.7696	0.7632
CGM	0.7144	0.6954	0.7289	0.7249	0.7296	0.6599	0.7103	0.6860
NB	0.6924	0.6649	0.7107	0.6971	0.6311	0.6266	0.6453	0.6472
Keyword	0.4044	0.3506	0.4109	0.4303	0.2644	0.2335	0.2668	0.2575
BFS	0.5368	0.3881	0.5129	0.5963	0.4521	0.6794	0.5090	0.6908
PageRank	0.5060	0.4843	0.5100	0.5225	0.5142	0.5008	0.5200	0.5425
VSM	0.2919	0.2147	0.3091	0.3089	0.1349	0.1272	0.1383	0.1476
HFN	0.1109	0.0904	0.1213	0.1174	0.2293	0.2204	0.2365	0.2451

site pages. This finding suggests that future work focusing on classifying the topics and sentiments at the page section level (e.g., forum page message-level) might be able to improve recall in forums for focused crawling tasks involving opinions and sentiments.

4.6. Subtopic Experiment

In the main experiment, the content of interest was negative sentiments towards drugs (PMDS) and companies thought to infringe on animal rights (CSR). However, various stakeholder groups may be interested in sentiments at a finer level of granularity (e.g., sentiments about a particular category of drugs, or within a certain industry). Accordingly, we evaluated the effectiveness of GBS and the comparison methods on six subtopics within the PMDS and CSR test beds. From PMDS, the four topics evaluated were drug categories: anti-viral drugs, anti-biotics, psychiatric drugs, and statins. In all four cases, pages containing negative sentiments towards these drug categories

Table VI. Subtopic Test Bed Statistics

Topic	Gold Standard Classifier Evaluation			Test Bed Statistics	
	Overall Accuracy	Relevant Recall	Irrelevant Recall	Relevant Pages	Irrelevant Pages
CSR-Restaurant	97.2 (2000)	95.2 (231)	97.5 (1769)	12,133	512,205
CSR-Apparel	96.3 (2000)	96.3 (214)	96.3 (1786)	11,357	512,981
PMDS-Anti-Viral	95.5 (2000)	92.7 (205)	95.8 (1795)	101,305	12,261,101
PMDS-Anti-Biotic	96.3 (2000)	96.9 (191)	96.2 (1809)	129,192	12,233,214
PMDS-Psychiatric	94.3 (2000)	95.2 (166)	94.2 (1834)	169,786	12,192,620
PMDS-Statin	95.8 (2000)	96.1 (181)	95.8 (1819)	99,202	12,263,204

were considered relevant. From CSR, two topics were selected: the restaurant and apparel industries. Pages containing negative sentiments regarding these two topics were considered relevant. In order to generate a gold standard, for each subtopic, we trained a gold standard SVM classifier using 800 relevant and 800 irrelevant pages per subtopic. As with the previous experiment, these training pages were identified by the four domain experts. Also consistent with the previous experiment, the SVMs were each trained using over 10,000 attributes, including word n-grams, parts-of-speech tag n-grams, as well as various lexical and syntactic measures [Abbasi and Chen 2008], learned from their respective 1,600 Web page training sets. These SVM classifiers were applied to the entire test beds to generate the gold standards. Each SVM was evaluated on a 2,000 page test set developed by the domain experts.

Table VI shows summary statistics for each of the six subtopics. The subtopic-level gold standard classifiers attained higher accuracy and recall rates than the general topic-level classifiers, when evaluated on 2,000 test pages (with overall accuracies between 94% and 97%). In other words, having higher topical specificity seemed to improve the gold standard classifiers' recall rates. Consistent with the previous experiment, the SVMs each took approximately 0.5 seconds per page, making them well-suited for gold standard construction but impractical for focused crawling. Looking at the right side of Table VI, the number of relevant pages per topic was considerably lower than for the two main tasks. Figure 10 shows the percentage breakdown of relevant pages across levels for each subtopic on the PMDS (left) and CSR (right) test beds. The values represent the averages across the 10 bootstrap runs, based on the shortest paths between seed URLs and test bed pages. The y-axis denotes the percentage of total relevant pages found at that particular level. Based on the left chart, the relevant pages associated with the anti-viral and anti-biotic drug categories tended to be concentrated in out-link levels 3–5 of the collection, while relevant psychiatric and statin drug pages occurred mostly in levels 3 and 4. With respect to the CSR subtopics, relevant restaurant pages were spread out across out-link levels 2–5 while target apparel industry content was mostly found in levels 3 and 4. As discussed below, this diversity in the content and occurrence patterns across subtopics had important implications for the performance of various crawlers.

GBS and the seven comparison methods were all evaluated on the six subtopic tasks. GBS, CGM, NB, Keyword, VSM, and HFN were all trained on the 1,600 training pages associated with each subtopic. All methods were once again evaluated using 10 bootstrap runs. For each subtopic, in each bootstrap run 200 seed URLs were randomly selected from 500 potential relevant URLs. Consistent with the previous experiment, there was no overlap between the 1,600 training pages, collection starting URLs used to construct the CSR and PMDS test beds, and crawl seed URLs used in the 10 bootstrap runs for each of the six subtopics. The experiment results for recall are presented in Figure 11. Based on the figure, GBS had the highest recall rates throughout

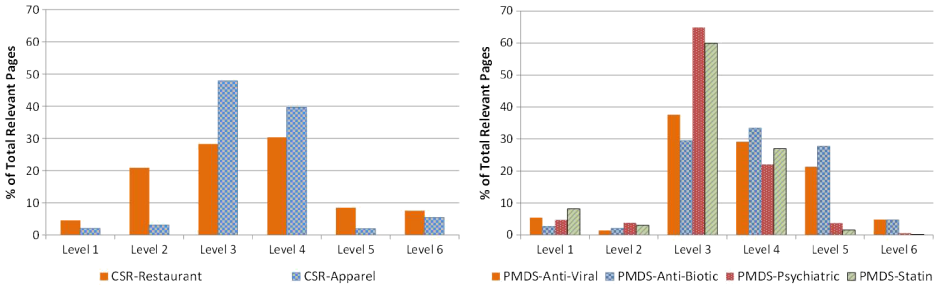

Fig. 10. Percentage of total relevant subtopic pages at each out-link level.

Table VII. Standardized Area Under the Curve (AUC) Values for Recall

Technique	CSR Test Bed		PMDS Test Bed			
	Restaurant	Apparel	Anti-Viral	Anti-Biotic	Psychiatric	Statin
GBS	0.7605	0.8065	0.7710	0.7926	0.7234	0.7661
CGM	0.6954	0.6791	0.7193	0.7341	0.6676	0.6754
NB	0.6720	0.6525	0.6042	0.6559	0.6249	0.6254
Keyword	0.4349	0.3521	0.3002	0.3199	0.2313	0.2494
BFS	0.5355	0.4625	0.5024	0.4469	0.6702	0.6749
PageRank	0.5066	0.4945	0.5648	0.5782	0.5247	0.5000
VSM	0.3229	0.2367	0.1595	0.1690	0.1254	0.1241
HFN	0.1183	0.0904	0.2674	0.2795	0.2190	0.2486

the crawl for each subtopic. With respect to the comparison methods, on the anti-viral and anti-biotic subtopics, CGM had the second best performance followed by NB. However, BFS performed well on the psychiatric and statin subtopics, outperforming CGM and NB. This was largely attributable to the high concentration of relevant pages at out-link level 3 for these two subtopics which allowed BFS to perform well during the first half of the crawl. With respect to the restaurant and apparel subtopics, after GBS, CGM, NB, and Keyword had the best performance.

Table VII shows the standardized AUC values for recall. GBS had AUC values that were 5%−12% higher than the next best method. CGM and NB outperformed the rest of the comparison methods. The improved performance of NB over CGM underscores the positive impact of incorporating tunneling in the crawler. The subtopic recall AUC values for GBS were comparable to those attained for the two main tasks. This suggests that while the ratio of relevant to irrelevant pages was lower for the subtopic tasks, this effect was counterbalanced by the higher topical specificity, resulting in comparable overall recall rates for GBS. In contrast, nonfocused crawlers such as BFS were adversely impacted by the lower ratio of relevant to irrelevant pages, resulting in lower recall AUC values on the subtopic tasks.

4.7. Impact of Sentiment Information and Graph-Based Tunneling

The experimental results presented in Sections 4.4 through 4.6 demonstrate the effectiveness of the GBS crawler in sentiment-driven crawling tasks. We conducted further analysis to understand the individual contribution of two important elements of GBS: the sentiment classifier and graph-based tunneling mechanism. We performed ablation analysis where GBS was compared against two variations. The first was GBS without tunneling (GBS-T), in which the graph comparison module was not utilized. GBS-T only relied on the text classifier (described in Section 3.1) to assign relevance

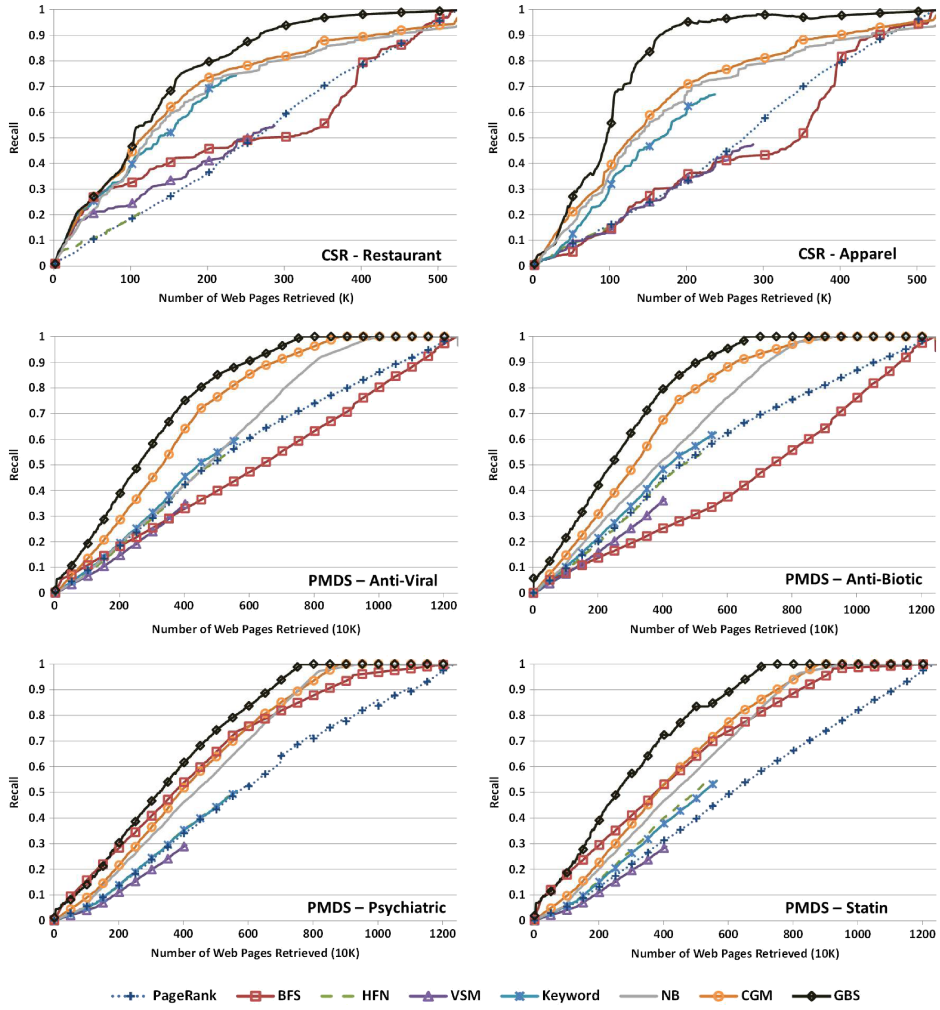

Fig. 11. Recall trends for subtopic pages in CSR and PMDS test beds.

weights to pages. C2, C3, and C4 pages (i.e., those deemed irrelevant by the text classifier) were never moved up in the queue since there was no tunneling mechanism. The other variation was GBS without tunneling or sentiment information (GBS-TS). Like a traditional topical crawler, GBS-TS weighted all pages purely on the basis of topical relevance, using the topic classifier described in Section 3.1. The comparisons between GBS, GBS-T, and GBS-TS were intended to isolate the impacts of the labeled web graph based tunneling module and the sentiment classifier, respectively. The comparison between GBS and GBS-TS was designed to illustrate the collective impact of the tunneling module and sentiment classifier.

Figure 12 shows the F-measure, precision, and recall trends for GBS, GBS-T, and GBS-TS on the CSR and PMDS test beds. On CSR, all three settings performed comparably over the first 20k pages. However, GBS performed considerably better for the remainder of the crawl. On PMDS, GBS performed better than GBS-T and GBS-TS from the onset. The difference in performance between GBS and GBS-T, which was quite noticeable on both test beds, demonstrates the usefulness of the graph-based

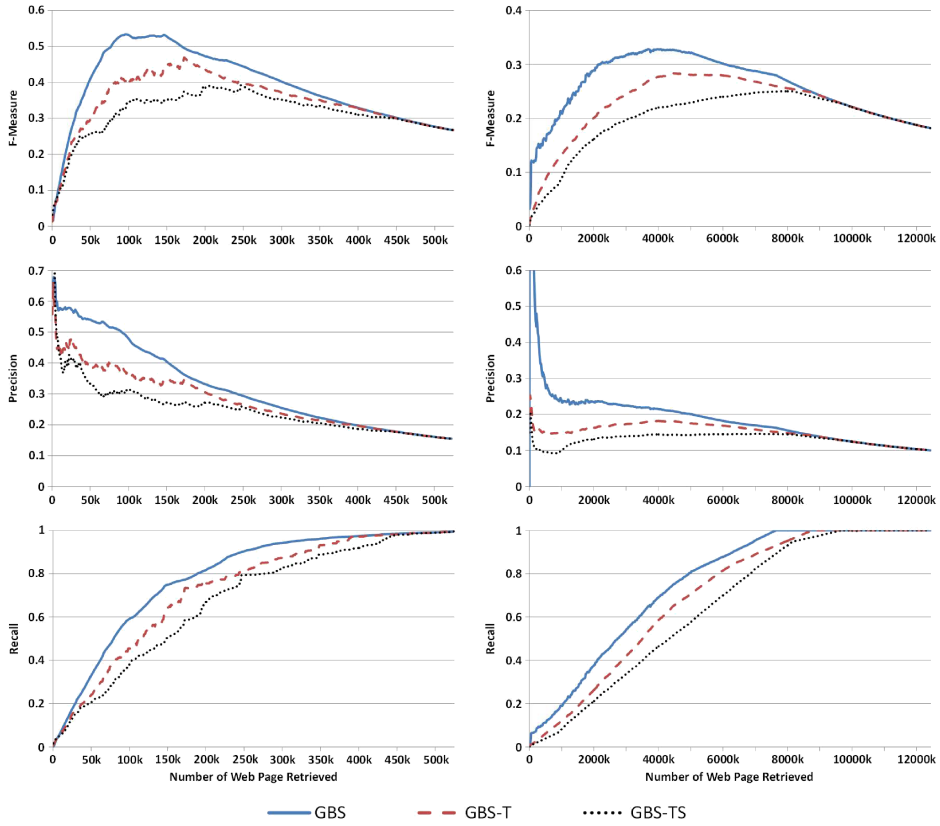

Fig. 12. F-Measure, recall, and precision trends for GBS, GBS-T, and GBS-TS on the CSR (left) and PMDS (right) test beds.

tunneling mechanism. Similarly, the performance gain yielded by GBS-T over GBS-TS illustrates the utility of the sentiment classifier employed by GBS. The results suggest that the tunneling and sentiment modules of GBS contributed significantly to the method's overall effectiveness. Moreover, given the more pronounced performance gap on the PMDS test bed, tunneling and sentiment information may be even more important during the crawl process when performing larger-scale collection of opinionated content. Collectively, the results presented in Figure 12 underscore the effectiveness of two critical components of the GBS crawler.

In addition to improving collection precision and recall, GBS was designed to run in a computationally efficient manner. By using random walk path based inexact graph matching, the graph-based tunneling module incorporated by GBS was able to evaluate pages in a computationally efficient manner. The tunneling mechanism had an average run time of 14.5 milliseconds per candidate page evaluated. The GBS crawler as a whole had an average crawl rate of 50 pages per second when run on a standard workstation.

5. CONCLUSIONS

In this study, we proposed GBS, a focused crawler that uses a graph-based tunneling mechanism and a text classifier that utilizes topic and sentiment information. Two major contributions of our study are as follows. First, we demonstrated that sentiment

information is useful for crawling tasks that involve consideration of content encompassing opinions about a particular topic. Second, we presented a novel graph-based method that ranks links associated with pages deemed irrelevant by utilizing labeled web graphs composed of nodes labeled with topic and sentiment information. This method helped GBS learn tunneling strategies for situations where relevant pages were near irrelevant ones. Collectively, these elements allowed GBS to outperform seven comparison crawling methods in terms of F-measure, precision, and recall on two test beds. For the majority of the crawls, GBS had recall rates that were at least 10% higher than the best comparison method. Moreover, GBS attained better recall rates at virtually all six levels on both test beds. Furthermore, GBS performed better on various categories of Web 2.0 content, including blog, forum, and social networking pages. GBS also outperformed comparison methods on six subtopics within the two test beds, demonstrating the effectiveness of the method for tasks involving different levels of information specificity. The experimental results suggest that GBS is able to collect a large proportion of relevant content after traversing fewer pages than existing focused crawlers. Additionally, the graph-based tunneling module utilized by GBS is computationally efficient, making it suitable for "real-time" data collection and analysis. Overall, the findings support the notion that focused crawlers that incorporate sentiment information are well suited to support Web 2.0 business and marketing intelligence gathering efforts.

REFERENCES

ABBASI, A. AND CHEN, H. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst. 26*, 2, Article 2.

ABBASI, A., CHEN, H., AND SALEM, A. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst. 26*, 3, Article 12.

ABBASI, A., FRANCE, S. L., ZHANG, Z., AND CHEN, H. 2011. Selecting attributes for sentiment classification using feature relation networks. *IEEE Trans. Knowl. Data Engin. 23*, 3, 447–462.

AGGARWAL, C. C., AL-GARAWI, F., AND YU, P. S. 2001. Intelligent crawling on the World Wide Web with arbitrary predicates. In *Proceedings of the 10th International Conference on World Wide Web*. 96–105.

ALLWEIN, E. L., SCHAPIRE, R. E., AND SINGER, Y. 2001. Reducing multiclass to binary: A unifying approach for margin classifiers. *J. Mach. Learn. Res. 1*, 113–141.

ARASU, A., CHO, J., GARCIA-MOLINA, H., PAEPCKE, A., AND RAGHAVAN, S. 2001. Searching the Web. *ACM Trans. Intern. Techn. 1*, 1, 2–43.

BAEZA-YATES, R. 2000. An image similarity measure based on graph matching. In *Proceedings of the 7th International Symposium on String Processing and Information Retrieval*. 28–38.

BHATTACHARYA, C., KORSCHUN, D., AND SEN, S. 2009. Strengthening stakeholder-company relationships through mutually beneficial corporate social responsibility initiatives. *J. Business Ethics 85*, 2, 257–272.

BREWER, T. AND COLDITZ, G. A. 1999. Postmarketing surveillance and adverse drug reactions. *J. Amer. Med. Assoc. 281*, 9, 830–834.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst. 30*, 1–7, 107–117.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A., AND WIENER, J. 2000. Graph structure in the Web. *Comput. Netw. 33*, 1–6, 309–320.

CHAKRABARTI, S., VAN DEN BERG, M., AND DOM, B. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Comput. Netw. 31*, 11–16, 1623–1640.

CHAU, M. AND CHEN, H. 2003. Comparison of three vertical spiders. *IEEE Comput. 36*, 5, 56–62.

CHAU, M. AND CHEN, H. 2007. Incorporating web analysis into neural networks: An example in Hopfield Net searching. *IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 37*, 3, 352–358.

CHEN, H. 2009. AI, e-government, and politics 2.0. *IEEE Intell. Syst. 24*, 5, 64–86.

CHEN, H. AND ZIMBRA, D. 2010. AI and opinion mining. *IEEE Intell. Syst. 25*, 3, 74–80.

CHO, J. AND GARCIA-MOLINA, H. 2003. Estimating frequency of change. *ACM Trans. Intern Techn. 3*, 3, 256–290.

CHO, J., GARCIA-MOLINA, H., AND PAGE, L. 1998. Efficient crawling through URL ordering. In *Proceedings of the 7th World Wide Web Conference*.

CHUNG, K., DERDENGER, T., AND SRINIVASAN, K. 2011. Economic value of celebrity endorsement: Tiger Woods' impact on sales of Nike golf balls. CMU Working Paper. http://www.andrew.cmu.edu/user/derdenge/TWExecutiveSummary.pdf.

CONTE, D., FOGGIA, P., SANSONE, C., AND VENTO, M. 2004. Thirty years of graph matching in pattern recognition. *Int. J. Pattern Recog. Artif. Intell. 18*, 3, 265–298.

DAVISON, B. D. 2000. Topical locality in the Web. 2000. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 272–279.

DILIGENTI, M., COETZEE, F., LAWRENCE, S., GILES, C. L., AND GORI, M. 2000. Focused crawling using context graphs. In *Proceedings of the 26th International Conference on Very Large Data Bases*. 527–534.

ESHERA, M. A. AND FU, K. S. 1984. A graph distance measure for image analysis. *IEEE Trans. Syst. Man Cybern. 14*, 3, 398–408.

ESULI, A. AND SEBASTIANI, F. 2006. SentiWordNet: A publicly available lexical resource for opinion mining, In *Proceedings of the 5th Conference on Language Resources and Evaluation*. 417–422.

FU, T., ABBASI, A., AND CHEN, H. 2010. A focused crawler for Dark Web forums. *J. Amer. Soc. Inf. Sci. Techn. 61*, 6, 1213–1231.

FÜRNKRANZ, J. 2002. Hyperlink ensembles: A case study in hypertext classification. *Inf. Fusion. 3*, 4, 299–312.

GAREY, M. AND JOHNSON, D. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co Ltd.

KASHIMA, H., TSUDA, K. AND INOKUCHI, A. 2003. Marginalized kernels between labeled graphs. In *Proceedings of the 20th International Conference on Machine Learning*. 321–328.

LEVENSHTEIN, V. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*, 8, 707–710.

LI, X., CHEN, H., ZHANG, Z., LI, J., AND NUNAMAKER, J. 2009. Managing knowledge in light of its evolution process: An empirical study on citation network–based patent classification. *J. Manage. Inf. Syst. 26*, 1, 129–153.

LIU, B. 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* 2nd Ed. Springer.

LIU, H., YU, P. S., AGARWAL, N., AND SUEL, T. 2010. Guest editors' introduction: Social computing in the Blogosphere. *IEEE Intern. Comput. 14*, 2, 12–14.

LU, H., CHEN, H., CHEN, T., HUNG, M., AND LI, S. 2010. Financial text mining: Supporting decision making using Web 2.0 content. *IEEE Intell. Syst. 25*, 1, 78–82.

MARTIN, E., MATTHIAS G., AND HANS-PETER, K. 2001. Focused web crawling: A generic framework for specifying the user interest and for adaptive crawling strategies. http://www.dbs.informatik.uni-muenchen.de/~ester/papers/VLDB2001.Submitted.pdf.

MENCZER, F., PANT, G., AND SRINIVASAN, P. 2004. Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Internet Technol. 4*, 4, 378–419.

MYERS, R., WILSON, R., AND HANCOCK, E. 2000. Bayesian graph edit distance. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 6, 628–635.

PANT, G. AND SRINIVASAN, P. 2005. Learning to crawl: Comparing classification schemes. *ACM Trans. Inf. Syst. 23*, 4, 430–462.

PANT, G. AND SRINIVASAN, P. 2009. Predicting web page status. *Inf. Syst. Resear. 21*, 2, 345–364.

RIECK, K., KRUEGER, T., BREFELD, U., AND MÜLLER, K. 2010. Approximate tree kernels. *J. Mach. Learn. Resear. 11*, 555–580.

RIESEN, K. AND BUNKE, H. 2010. Graph classification and clustering based on vector space embedding. In *Machine Perception and Artificial Intelligence,* World Scientific Publishing Company. 348.

ROBLES-KELLY, A. AND HANCOCK, E. R. 2005. Graph edit distance from spectral seriation. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 3, 365–378.

SALTON, G. AND MCGILL, M. J. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc. New York, NY, 400.

SCHAPIRE, R. E. AND SINGER, Y. 1999. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn. 37*, 3, 297–336.

SHANNON, C. E. 1948. A mathematical theory of communication. *Bell Syst. Techn. J. 27*, 4, 379–423.

SPANGLER, S., PROCTOR, L., AND CHEN, Y. 2008. Multi-Taxonomy: Determining perceived brand characteristics from web data. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 258–264.

SRINIVASAN, P., MENCZER, F., AND PANT, G. 2005. A general evaluation framework for topical crawlers. *Inf. Retrieval* 8, 417–447.

SUBRAHMANIAN, V. S. 2009. Mining online opinions. *Comput. 42*, 7, 88–90.

THELWALL, M. 2007. Blog searching: The first general-purpose source of retrospective public opinion in the social sciences? *Online Inf. Rev. 31*, 3, 277–289.

TREMAYNE, M., ZHENG, N., LEE, J. K., AND JEONG, J. 2006. Issue publics on the Web: Applying network theory to the war Blogosphere. *J. Comput.-Mediated Commun. 12*, 1, Article 15.

VAN GROOTHEEST, K., DE GRAAF, L., AND DE JONG-VAN DEN BERG, L. T. 2003. Consumer adverse drug reaction reporting: A new step in pharmacovigilance? *Drug Safety 26*, 3, 211–217.

WIEBE, J. M. 1994. Tracking point of view in narrative. *Comput. Linguistics 20*, 2, 233–287.

YANG, Y. AND PEDERSEN, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*. 412–420.

Received April 2011; revised March 2012; accepted June 2012