Improve Web Search Using Image Snippets

Xiao-Bing Xue and **Zhi-Hua Zhou**
National Laboratory for Novel Software Technology
Nanjing University
Nanjing 210093, China
{xuexb, zhouzh}@lamda.nju.edu.cn

Zhongfei (Mark) Zhang
Department of Computer Science
SUNY at Binghamton
Binghamton, NY 13902, USA
zhongfei@cs.binghamton.edu

Abstract

The Web has become the largest information repository over the world. Therefore, effectively and efficiently searching the Web becomes a key challenge. Previous research on Web search mainly attempts to exploit the text in the Web pages and the link information between the pages. This paper shows that the Web search performance can be enhanced if image information is considered. In detail, a new Web search framework is proposed, where *image snippets* are extracted for the Web pages, which are then provided along with text snippets to the user such that it is much easier and more accurate for the user to identify the Web pages he or she expects and to reformulate the initial query. Experimental evaluations demonstrate the promise of the proposed framework.

Introduction

With the explosive development of the Internet, the Web has become the largest information repository over the world. Due to its huge volume, users easily feel lost in this repository. Web search engines attempt to help users find a needle in the haystack. It is not surprising that Web search techniques have attracted more and more attention in the related communities.

In the literature, text information was the first to be considered to exploit for Web search, as Web pages contain abundant text. Many techniques from text retrieval were applied to Web search, such as Vector Space Model (VSM) and TFIDF (Salton & Buckley 1988). Later, exploiting link information began to receive substantial attention in the literature, noticeably such as Page Rank (Brin & Page 1998) and HITS (Kleinberg 1999).

In addition to *text* and *link*, other modalities of information, such as *imagery*, *video*, *audio*, also convey rich and important information about a Web page content. Yet, while there is substantial literature on multimedia information retrieval (Frankel & Swain 1996; Smith & Chang 1996), which focuses on retrieving imagery, video, audio, etc. from the Web, it is a regretful reality that there are no mature techniques to effectively exploit these modalities of information for Web search, which focuses on retrieving Web pages.

This paper focuses on the imagery modality to further assist the Web search. It is well known that automatic im-

age understanding is rather difficult in general, and it is specifically true for Web images with typically diverse visual quality and semantics while automatic text processing is more mature than image understanding in the context of Web search. On the other hand, it is also observed that human vision works better and faster for image understanding than for text processing. For example, it is reported that human can get the gist of an image in 110ms or less (Coltheart 1999), while in these 110ms, human can only read less than 1 word, or skim 2 words (The average English reader can read about 4.2 words per second, and can skim or scan at roughly 17 words per second (Chapman 1993)). This work is thus motivated to take advantage of this fact to bring imagery to the human interaction part of Web search to develop an effective and efficient search framework.

Text snippet is widely used in Web search. It is a summarization of the retrieved Web page, that helps a user identify the content of the page. Considering the complementarity between text and imagery, this paper proposes the similar *image snippet* concept for a Web page. Image snippet is one of the images in a Web page that is both representative of the theme of the page and at the same time closely related to the search query. A Web page may or may not necessarily have an image snippet as in an extreme case a Web page may not even have images at all. However, it is arguably obvious that image snippets do exist in many Web pages. Based on this consideration, a new Web search framework, WebSIS (Web Search using Image Snippets), is developed in this paper.

In the traditional Web search, the search process can be typically described as follows: first, a search query is posed by a user; second, the search result consisting of text snippets of the retrieved Web pages is returned to the user for browsing; third, if the search result is satisfactory to the user, the search is completed; if not, which is the typical case, the user may reformulate the query through labelling relevant/irrelevant pages or directly revising the query terms manually. The second and third steps may repeat until the final result becomes satisfactory to the user. This paper focuses on the user interaction part of the Web search; other parts such as Web page crawling, Web page cleaning, and Web page indexing are not considered. In WebSIS, image snippets are applied to the second and third steps.

For the second step, image snippets are provided along with text snippets to the user. Since human 'reads' images

much faster than text, it is easier for a user to locate the relevant image snippets than the relevant text snippets. These relevant image snippets indicate the potential Web pages required by the user. Then, since image snippets and text snippets provide somewhat complementary information to each other, they are combined to help the user identify the Web pages he/she expects from potential Web pages more accurately. Note that, since imagery can be perceived quickly by humans, the use of image snippets brings more information to the user about the content of a Web page at the cost of little additional cognitive load to the user. In contrast, if this more information is brought through using more detailed text snippets, the load to the user would be significantly increased.

For the third step, according to both text snippets and image snippets, a user selects relevant Web pages returned in the previous round of retrieval as the feedback to allow the search engine automatically reformulate the query for the next round of retrieval. As mentioned above, image snippets help the user identify relevant pages more accurately, and therefore, these more accurate labels also improve the performance of this kind of query reformulation. In case the user prefers revising the query manually, which usually occurs when the process of labelling relevant Web pages is boring and time-consuming, in the traditional search scenario, the user would have difficulty and challenge to generate more relevant terms for reformulating the query as otherwise he/she would have had these terms ready in the beginning. In WebSIS, this is not the case, as the returned image snippets are collected and displayed at the top of the result page for feedback; the user simply needs to select relevant image snippets as the feedback to allow the search engine automatically reformulate the query. Since the user 'reads' images faster than text, it takes much less time for this process than that for selecting relevant Web pages.

In general, there are two advantages of using image snippets. First, since human 'reads' images faster than text, image snippets help the user locate the potential Web pages he/she expects more quickly and the proposed query reformulation purely using image snippets for feedback costs the user little time while not requiring the user to design additional query terms manually. Second, since imagery could provide somewhat complementary information to text, providing image snippets along with text snippets helps the user find the Web page he/she truely expects more accurately. Consequently, the effect of the query reformulation using the retrieved Web pages for feedback is improved.

The contribution of this paper lies in: we have proposed the image snippet concept complementary to the existing text snippets in the traditional Web search; with image snippets, exploiting the advantage of human perception for imagery over for text and the complementarity between imagery and text, we have developed the novel Web search framework WebSIS to deliver more effective and efficient Web search.

The rest of this paper starts with a brief review on related works. Then, the WebSIS framework is presented and experimental evaluations are reported, which is followed by the conclusion.
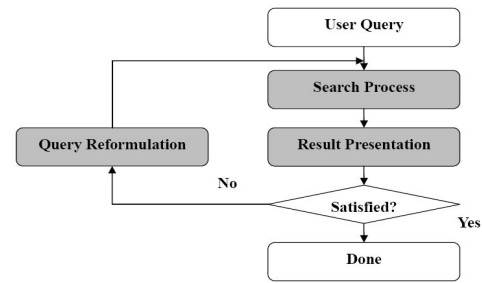

Figure 1: The WebSIS framework

Related Work

Query term matching and link structure analysis are the major techniques used by commercial Web search engines. Recently, several new techniques have been proposed to improve the Web search performance. Cai et al. (2004b) took a block of a Web page as a single semantic unit instead of the whole page and proposed the block-based Web search. This search framework solves two problems that typically documents may contain multiple drifting topics and may have varying lengths. Ntoulas et al. (2005) proposed a new Web search engine called *Infocious* which improves the Web search through linguistic analysis to resolve the content ambiguity of Web pages and to rate the quality of the retrieved Web page.

There is significant literature on multimedia information retrieval, especially image retrieval on the Web (Frankel & Swain 1996; Smith & Chang 1996; Cai *et al.* 2004a). However, little work has been done to exploit other modalities of information, especially visual information, to improve the Web search. Woodruff et al. (2001) had an attempt in this direction. In their work, an enhanced thumbnail was proposed to help the user search the Web. This enhanced thumbnail is an image, where the whole Web page is resized to fit into this image and important text is highlighted in order to be readable for the user. It is reported that with this enhanced thumbnail the user can find the answer in less time than with the text snippet.

In contrast, in other fields, imagery has played an important role in complementing text. Barnard and Johnson (2005) used images for word sense disambiguation. In their work, a statistical model for the joint probability of image regions and words is learned to automatically annotate images. The probabilities of all the possible senses of a word are compared and the most probable sense is determined. Better results are reported when this strategy is used to augment the traditional word sense disambiguation.

The WebSIS Framework

In this section, the proposed WebSIS framework is introduced in detail and the implementation is also discussed.

Framework

The framework of WebSIS is illustrated in Figure 1. The basic process of WebSIS is similar to that of the traditional Web search except that different operations are required in the grey boxes in Figure 1.

Table 1: Algorithm for image snippet extraction

ALGORITHM: Image Snippet Extraction
INPUT: Web page *Page*, search query *Query*
OUTPUT: image snippet *Image*
PROCESS:

1. Segment *Page* into blocks using a Web page segmenter and store the blocks in *B*.

2. Evaluate the importance of each block in *B* using a Web block evaluator and store important blocks in *ImpB*.

3. Determine the similarity between the *Query* and each block in *ImpB* that contains at least one image, and store those images appearing in the block with the highest similarity value in *CanImg*.

4. Select *Image* from *CanImg*

   - IF *CanImg* is empty, return NULL.
   - ELSE return *Image* according to the specified heuristic rule.

In *Search Process*, the major difference is that, for a given query, the image snippets are extracted from each Web page retrieved. As defined, an image snippet should be representative of the theme of the Web page, and at the same time also relevant to the query. In order to identify a representative image, a Web page segmenter is first used to segment the original Web page into several blocks, and then a Web block evaluator is used to further identify the important blocks. The images in important blocks are considered as potential representative images. In order to finally identify relevant images, a similarity between the query and all important blocks containing at least one image each is measured. The images appearing in the block with the highest similarity value are considered as both representative and relevant. If multiple images are identified at this point, heuristic rules are applied to break the tie to select one of the images. The algorithm for the image snippet extraction is summarized in Table 1.

With the image snippets, the *result page* is obtained, which is the basis for *Result Presentation* and *Query Reformulation*. Figure 2 illustrates the differences between a result page of WebSIS and that of the traditional Web search.

Part 1 of Figure 2 is the part used for *Result Presentation*. Clearly, in WebSIS, an image snippet is provided along with the corresponding text snippet. With the help of the image snippet, users can easily and accurately identify the Web pages they expect.

In *Query Reformulation*, two types of query reformulation are provided. *Type I* requires a user to label whether a retrieved Web page is relevant or not according to the provided snippets and then all these labels are collected as the feedback information to automatically reformulate the query. With the explicit help of image snippets, it is easier for the user to more accurately determine the relevancy of the retrieved Web pages to the query. Consequently, this

a) traditional Web search   b) WebSIS

Figure 2: Comparison of the result page

type of query reformulation is anticipated to be more effective. For *Type II*, a user simply selects some relevant images from the feedback pool as shown in Part 2 of Figure 2 to allow the system automatically reformulate the query. In contrast, in the traditional Web search, no such feedback pool is provided and users have to design additional query terms by themselves.

Implementation

While the WebSIS framework is a general approach to Web search, in this subsection, the implementation of a prototype of the WebSIS framework is discussed.

The major work of *Search Process* is to implement the algorithm of the image snippet extraction as described in Table 1. The discussions refer to the steps defined in the algorithm in Tabel 1.

For Step 1, many techniques can be used to segment a Web page. Here, the recently proposed VIPS (VIsion-based Page Segmentation) algorithm (Cai *et al.* 2003) is used as the Web page segmenter. The basic idea of this technique is to emulate how a user understands the Web page layout based on the user's visual perception. Usually, the visually closely packed regions are likely to share the same semantics. Such regions are thus extracted as a block of the page. A simple example is given in Figure 3[1]. The detailed description about the VIPS algorithm can be found in (Cai *et al.* 2003).

For Step 2, the Web block importance model proposed by (Song *et al.* 2004) is used as the Web block evaluator. In this model, two types of features, spatial features and content features, are extracted for each block. In the original model, importance of four different levels is assigned to each block in the training set by manually visual inspection. These instances are then fed to a neural network or SVM to learn the importance model. The detail of this model can be found in (Song *et al.* 2004). In WebSIS prototype, however, since each block is either important or not, a simplified, binary importance model is used.

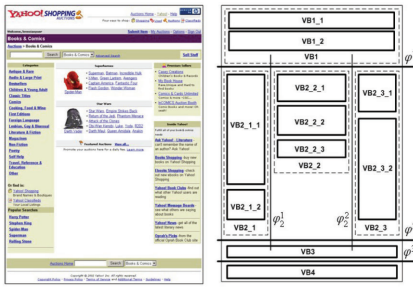[1]cited from http://www.ews.uiuc.edu/~dengcai2/VIPS/VIPS .html

Figure 3: A simple example of the VIPS algorithm.

For Step 3, the cosine function is used as the similarity function.

For Step 4, typical heuristic rules used to select one from a group of tied images include identifying the image with the largest size or identifying the image with its ALT field containing the query term. In the WebSIS prototype, the image appearing first in the block is selected as the image snippet.

The two types of *Query Reformulation* are implemented in the prototype separately and are discussed below.

For *Type I*, the standard relevance feedback technique in text retrieval is used to automatically reformulate the query. Each term in the relevant Web pages except the query item is selected as a candidate. Then, the *wpq* method (Robertson 1990) is used for ranking all the candidates. The weight of a candidate term t is determined in Eq. 1, where r_t is the number of the seen relevant documents containing term t, n_t is the total number of the documents containing t, R is the number of the seen relevant documents for query q, and N is the total number of the documents in the collection. After all the candidates are sorted according to *wpq* score, the *ExpNum* (the number of the terms used for query expansion) terms with the largest *wpq* score are added to the original query.

$$wpq_t = \log \frac{r_t/(R - r_t)}{(n_t - r_t)/(N - n_t - R + r_t)} \times (\frac{r_t}{R} - \frac{n_t - r_t}{N - R}) \tag{1}$$

For *Type II*, the implementation is described as follows. First, all images in the important blocks are used as candidates. Second, for each candidate image, its text description consists of three parts: the title of the Web page, the ALT field of this image, and the text in the block where this image appears. Given the text description, the similarity between the query and the candidate images can be immediately determined. The similarity values are then used to sort the candidate images. Finally, the first *PoolNum* (the number of the images put into the feedback pool) images are selected to be put into the feedback pool. Given the feedback of the user, the relevance feedback technique used in *Type I* is also used here, except that the candidate terms come from the text description of the image, but not from the whole Web page.

Experiments

In this section, experiments are reported for *Result Presentation* and *Query Reformulation*, respectively.

Table 2: Three types of queries

Type	Query
ambiguous	tiger, apple, dove, eagle, jaguar, jordan, newton, aaai, cambridge, trec
unambiguous	tiger beer, apple fruit, dove chocolate, eagle bird, jaguar car, jordan basketball, newton physics, aaai artificial intelligence, cambridge university, trec information retrieval
hot	digital camera, hurricane katrina, ares, brad pitt, ipod nano, janet jackson, myspace, orkut, xbox

Table 3: The contingency table

		Label of Web page	
		Yes	No
Label of	Yes	TP	FP
Snippet	No	FN	TN

Experimental Configuration

In order to design realistic Web search experiments, three types of queries are designed: *ambiguous query*, *unambiguous query* and *hot query*. 10 queries are respectively designed for the ambiguous query and the unambiguous query categories. 9 queries chosen from Google 2005 top searches[2] are used for the hot query category. The details of these queries are documented in Table 2. Note the one-to-one correspondence between the queries in the ambiguous category and the unambiguous category. For example, the query 'jaguar' in the ambiguous category and the query 'jaguar car' in the unambiguous category correspond to the same information requirement. A passage of text is provided for each query, which describes the information requirement of this query in detail.

For each query, the first 200 Web pages returned by Google search engine and their corresponding text snippets are downloaded. These 200 Web pages consist of a small corpus and experiments for each query are conducted on this corpus. The search strategy used here is the traditional Vector Space Model and TFIDF weighting method. Since experiments are mainly designed to test the promise and effectiveness of WebSIS for user interaction, 9 volunteers are involved in experiments. For each query, experiments are conducted with three different users. The relevant Web pages for each query is obtained through letting the volunteers examine all 200 Web pages. Note that the volunteers examine 200 Web pages after they have completed all the experiments described as follows, which guarantees that this process does not influence the validity of the subsequent experiments. For image snippet extraction, equal rank is given to those important blocks with at least one image in the following experiments for simplicity. Among those blocks, the one given the largest order by the VIPS algorithm is simply selected for extracting the image snippet.

---

[2]http://www.google.com/intl/en/press/zeitgeist2005.html

Table 4: The result of different snippets

	p	r	$F1$	acc
Image	0.637	0.515	0.569	0.764
Text	0.607	0.693	0.647	0.797
Text+Image	0.656	0.753	0.701	0.842

Table 5: The time spend on labelling different snippets

	Image	Text
Time(second)	2.34	7.02

Experiments on Result Presentation

This series of experiments are designed to test the helpfulness of introducing image snippets to *Result Presentation*. For each query, initially only the image snippets are presented to a volunteer and he/she is asked to label these snippets as relevant or irrelevant. The time spent is recorded. Then, only the text snippets are presented and the same procedure is repeated. The time spent is also recorded. Finally, the text snippets and the image snippets are both presented and the volunteer is asked to label the combination of image snippets and text snippets. In each stage, the order of snippets presented to the user is randomly shuffled. Since this experiment requires that each Web page have an image snippet, it is only conducted for the Web pages where the image snippet is available.

The standard precision, recall, F1 measure and accuracy are used as the performance measure. Given the contingency table in Table 3, precision (p), recall (r), F1 measure ($F1$), and accuracy (acc) can be determined as follows. The average result of 87 (29×3) user-query combinations is shown in Table 4. The average time spent on labelling snippets is shown in Table 5.

$$p = \frac{TP}{TP + FP} \qquad r = \frac{TP}{TP + FN}$$
$$F1 = \frac{2 \times p \times r}{p + r} \qquad acc = \frac{TP + TN}{TP + FP + FN + TN}$$

Table 4 shows that only using text snippet is better than only using the image snippet. However, as expected, when the text snippet and the image snippet are combined simultaneously, a result better than that of each component only is obtained, which verifies that presenting both the text snippet and the image snippet helps users identify the Web pages they expect more accurately.

Figure 4 illustrates the complementarity of the text and image snippets. The two retrievals are both about the query 'ipod nano' and the information requirement for this query is 'any Web page that can help you buy the ipod nano mp3 player'. In Figure 4(a), '200GB ipod nano' in the text snippet is sufficient to let most users believe that this Web page introduces this mp3 player and should be relevant (users familiar with ipod nano may feel strange about the huge disk volume). When the image snippet is given, things become clearer. This Web page is about how to connect the mp3 player with a hard disk and that is why the mp3 player has 200GB volume. Obviously, this Web page is irrelevant. In Figure 4(b), the image snippet convinces most users that this Web page is relevant, while the text snippet tells users that
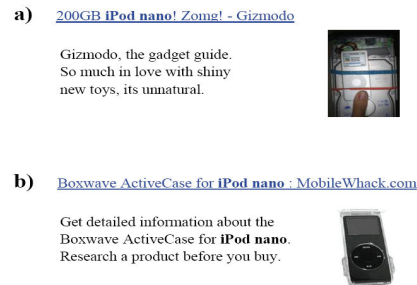

a) 200GB **iPod nano**! Zomg! - Gizmodo

Gizmodo, the gadget guide. So much in love with shiny new toys, its unnatural.

b) Boxwave ActiveCase for **iPod nano** : MobileWhack.com

Get detailed information about the Boxwave ActiveCase for **iPod nano**. Research a product before you buy.

Figure 4: Examples of combing text and image snippets.

this Web page introduces the plastic case used for protecting the ipod nano.

Table 5 shows that labelling the text snippet is about 3 times slower than labelling the image snippet. This verifies that, with the help of the image snippet, users can locate the relevant Web page more quickly.

Experiments on Query Reformulation

For *Type I*, the effects of the image snippet only, the text snippet only, and the combination of both are compared when used for relevance feedback. Since in experiments on *Result Presentation*, the volunteers have labelled the image snippets, the text snippets, and the combination of both, these labels can be used in this experiments. Specifically, the labels for the first *FeedbackNum* (the number of the Web pages used for feedback) retrieved Web pages are used to automatically expand the original query. In the practical search environment, the user usually label a Web page only according to the snippet, thus in this experiment the labels of the image snippets, the text snippets, and the combination of both are respectively used as the labels of the retrieved Web pages. *ExpNum* is set to 10. The expanded query is posed again, and the retrieved Web pages based on the expanded query are evaluated. Here, P@10, the precision of the first 10 Web pages, is used as the performance measure. The average result with respect to *FeedbackNum* is depicted in Figure 5. The result of a baseline which is the performance of the initial query is also given. Since this experiment also requires that each Web page have an image snippet, it is only conducted for the Web pages where the image snippets are available.

Figure 5 discloses that the combination of the text snippet and the image snippet performs better than each component only.

For *Type II*, an experiment is designed to compare the effect of using relevant images in the feedback pool to reformulate the query and the effect of designing additional query terms manually by a user. Specifically, for each query, the volunteers first select relevant images from the feedback pool. These images are used to expand the original query and the performance of the expanded query is evaluated. Here, the performance measure is still P@10. Second, the volunteers design several additional query terms manually and the performance of this revised query is also evaluated. The average result with respect to *PoolNum* is shown in Figure 6. Note that since it is not necessary that an image snippet is always available in each Web page in this experiment,
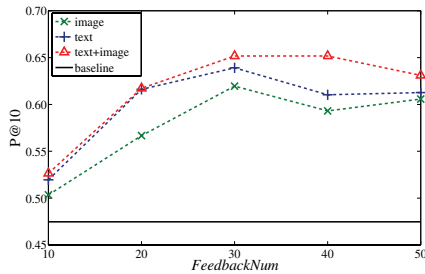
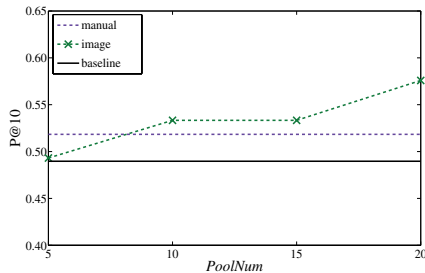Figure 5: Result of *Type I* with respect to *FeedbackNum*.

Figure 6: Result of *Type II* with respect to *PoolNum*.

this experiment is conducted for all the Web pages.

Figure 6 shows that when the number of images in the feedback pool is small (say, 5), the user designed query performs better than using image snippets for feedback, while with a larger number of images in the feedback pool, queries reformulated by purely using image snippets are even better than the user designed queries apparently.

Conclusion

In the previous research on Web search, the information exploited was mainly the text in the Web pages and the link between the pages. While there is substantial literature on multimedia information retrieval, especially image retrieval, little work strictly exploited image information in Web search. In this paper, a new Web search framework called WebSIS is proposed, where the image information is used to help improve the performance of Web search. Specifically, the concept of *image snippet* is proposed for Web pages and is used for the presentation of search results and the reformulation of the queries. Experiments show that image snippet can help users identify the Web pages they expect and reformulate the search query more effectively and efficiently.

It is noteworthy that the desirable 'efficiency' in Web search involves two aspects. First, the search process should be as quick as possible, which has been studied by many researchers. Second, the search process should enable the user to identify what he or she expects quickly, which is far from being well-studied. This paper attempts to tackle the second issue by exploiting image snippets. It is anticipated that a powerful Web search engine can be developed if techniques for addressing both the issues are combined gracefully.

Acknowledgements

References

Barnard, K., and Johnson, M. 2005. Word sense disambiguation with pictures. *Aritificial Intelligence* 167(12):13–30.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks* 30(1-7):107–117.

Cai, D.; Yu, S.-P.; Wen, J.-R.; and Ma, W.-Y. 2003. VIPS: A vision-based page segmentation algorithm. Technical report, No. MSR-TR-2003-79, Microsoft.

Cai, D.; He, X.-F.; Li, Z.-W.; Ma, W.-Y.; and Wen, J.-R. 2004a. Hierarchical clustering of WWW image search results using visual, textual and link analysis. In *Proceeding of the 12th ACM International Conference on Multimedia*, 952–959. New York, NY: ACM Press.

Cai, D.; Yu, S.-P.; Wen, J.-R.; and Ma, W.-Y. 2004b. Block based web search. In *Proceeding of the 27th ACM International Conference on Research and Development in Information Retrieval*, 456–463. New York, NY: ACM Press.

Chapman, A., ed. 1993. *Making Sense: Teaching Critical Reading Across the Curriculum*. New York, NY: The College Board.

Coltheart, V., ed. 1999. *Fleeting Memories: Cognition of Brief Visual Stimuli*. Cambridge, MA: MIT Press.

Frankel, C., and Swain, M. J. 1996. Webseer: An image search engine for the world wide web. Technical report, No. 96-1, University of Chicago, Computer Science Department, Chicago, IL.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5):604–632.

Ntoulas, A.; Chao, G.; and Cho, J. 2005. The infocious web search engine: Improving web searching through linguistic analysis. In *Proceeding of the 14th World Wide Web conference*, 840–849. New York, NY: ACM Press.

Robertson, S. E. 1990. On term selection for query expansion. *Journal of Documentation* 46(4):359–364.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5):513–523.

Smith, J. R., and Chang, S.-F. 1996. Searching for image and videos on the world wide web. Technical report, No. 459-96-25, Center for Telecoumunication Research, Columbia University, New York, NY.

Song, R.-H.; Liu, H.-F.; Wen, J.-R.; and Ma, W.-Y. 2004. Learning block importance models for web pages. In *Proceeding of the 13th World Wide Web conference*, 203–211. New York, NY: ACM Press.

Woodruff, A.; Faulring, A.; Rosenholtz, R.; Morrison, J.; and Pirolli, P. 2001. Using thumbnails to search the web. In *Proceeding of the SIGCHI Conference on Human Factors in Computing Systems*, 198–205. New York, NY: ACM Press.