Contextual Information Portals

Jay Chen, Trishank Karthik, Lakshminaryanan Subramanian
jchen@cs.nyu.edu, trishank.karthik@nyu.edu, lakshmi@cs.nyu.edu

There is a wealth of information on the Web about any number of topics. Many communities in developing regions are often interested in information relating to specific topics. For example, health workers are interested in specific medical information regarding epidemic diseases in their region while teachers and students are interested in educational information relating to their curriculum. This paper presents the design of *Contextual Information Portals*, searchable information portals that contain a vertical slice of the Web about arbitrary topics tailored to a specific context. Contextual portals are particularly useful for communities that lack Internet or Web access or in regions with very poor network connectivity. This paper outlines the design space for constructing contextual information portals and describes the key technical challenges involved. We have implemented a proof-of-concept of our ideas, and performed an initial evaluation on a variety of topics relating to epidemiology, agriculture, and education.

Introduction

The World Wide Web has completely transformed the way people interact with information. In particular, the ubiquitous availability, comprehensive content, and usability of the Web makes it the definitive resource for answers to questions and information in general. In the developed world a user with good Internet connectivity has the luxury of issuing multiple search queries to a search engine to retrieve the specific information of interest from the Web. In contrast, a large fraction of users in the developing world do not have this luxury because Web access has largely been confined to urban areas. In rural areas the Internet is unusable for a variety of reasons (Brewer et al. 2006) resulting in poor and unreliable connectivity. Unfortunately, providing reliable high bandwidth connectivity in these regions has not been economically viable due to a variety of factors including low purchasing power, low aggregate demand and the relatively high cost or lack of good backhaul connectivity in many countries. As a result, the relative price of connectivity is very high in developing regions (Mubaraq et al. 2005). Even where reasonably good bandwidth is available (such as a 1 Mbps leased line in universities and small companies etc), the network connection is shared across many users ($50 - 1000$ users in universities) resulting in poor per-user bandwidth. To make matters worse, the quality of Web pages have significantly advanced over the past few years; the dramatic increase in the size and complexity of Web pages causes a proportional increase in page rendering times on low-bandwidth connections.

This paper describes the design of a *Contextual Information Portal (CIP)*, a system that provides an *offline searchable and browseable portals composed of vertical slices of the Web about specific topics*. For a single topic, say "malaria", the goal of a CIP is to make available a large portion of relevant Web pages pertaining to the topic with little or no Internet access. Constructing this set of Web pages involves first crawling the Web for pages that are deemed useful, indexing and re-ranking them locally, storing the information on large storage media (e.g. hard disks or DVDs), and finally shipping the self-contained Web cache to its destination. We envision even small grassroots organizations or NGOs with comparatively few computing resources determining and building their own portals on topics specific to their user interests. Depending on the situation, a CIP may be integrated with the existing infrastructure to bootstrap the local cache or as a standalone portal within a kiosk service. As a standalone portal a CIP provides an interactive search and browsing interface enabling a Web-like experience for the topics covered. While building a complete CIP will involve dividing resources between topics based on their relative importance, in this paper we focus on the problem for an individual topic.

The basic premise of the CIP design is that the context informs the specific information needs of different communities of users across localities in developing regions. CIPs are designed primarily for environments where connectivity is very poor or not available. As a result, CIPs are a departure from the standard Web caching model of loading a large cache with the most popular Web pages. For these poorly connected locales the standard model has several problems: First, in our deployments in a large university in India, we have observed very low cache hit rates of less than 20% (Chen, Subramanian, and Li 2009). These frequent cache misses result in very slow and fluctuating page rendering times. Second, the interface to a proxy cache returns only a binary yes/no result of whether a specific object

is in the cache. This is problematic because it is highly possible that even if a specific page does not exist in the local cache a different but equally useful page does. This disjunction between what is indexed by large search engines and what is easily accessible by the user increases as a function of several variables which in aggregate are growing worse: the slowness of the connection, the increasing size of the average Web page, and the huge amount of pages indexed by large search engines. In this paper, we first describe a few example scenarios where CIPs can be useful. We go on to analyze several obvious but problematic approaches to constructing contextual portals, and where artificial intelligence techniques may be applied to address these problems. We then outline the process of building a CIP, and conclude by discussing some of our initial results and experiences.

CIP Example Scenarios

There are several real-world examples where contextual information portals are useful.

Agricultural portals: eChoupal (ech) is a large initiative by ITC in India that has established nearly $7,000$ kiosks with VSAT Internet connectivity in villages throughout India to directly link with farmers for procurement of agricultural produce especially soybeans and wheat. While the current usage model of these kiosks is limited, an India-specific information portal on soybeans and wheat could be an important value-added service that eChoupal can offer for its rural farmers at each kiosk. We have begun work with agricultural universities in India to develop such portals for *organic farming* and *water harvesting* practices.

Medical portals: Bluetrunk Libraries (blu) is a massive project by WHO to ship mobile libraries of 150 healthcare books into remote regions of Africa as an educational guide for healthworkers on the field. With the wealth of medical information online, one can envision a similar offline searchable medical portal for healthworkers. St. Johns Medical College, one of the large hospitals in India has expressed interest in a similar medical portal. The same idea can be extended to disease specific portals for important diseases such as HIV, TB, malaria, diabetes or for portals for specific specializations such as opthalmology or surgery. WiSE-MD (wis) is one specific example of a portal for surgical education that has adopted as a standard teaching module in many medical universities within the US.

Education portals: Given the volume of educational material available online, educational portals for specific topics may be automatically constructed separately or as supplementary materials to well structured portals such as MIT OpenCourseWare (ope). Portals for specific educational topics such as operating systems or artificial intelligence could be useful also in the developed world as domain specific digital libraries.

Locality specific web portals: Collaborative caching over delay tolerant networks (Isaacman and Martonosi) has been observed to improve cache hit rates dramatically due to similar interests across nearby villages. The prefetching component in these systems could be improved by learning which location specific topics to prefetch and when time varying topics should be updated (e.g. weather should be updated frequently).

Design

There are many different ways to construct an information portal. Three approaches of increasing sophistication are: (a) Index an existing Web cache; (b) Crawl the entire Web and extract only the pages relevant to the specified topic; (c) Issue several queries relating to the topic to a search engine and download the top results for each query.

In the first approach, if the existing cache is from another locale then it may not have the content that the target locale is interested in. Even in the degenerate case where the cache is from the target locale, it is tempting to conclude that the content is context specific, but what exists in the cache simply reflects the information that was requested, but does not indicate whether it was actually useful. This inherent ambiguity of deriving implicit user intent from cache contents and/or access logs means we some thought must go into deriving topics from logs. For example, there is a subtle difference between the page requests themselves and the topics in which the pages belong.

The second approach is more comprehensive, but extremely expensive in terms of the number of wasteful pages that need to be downloaded before finding an appropriate set of pages to include in the CIP. It is possible for a large search engine which already has a local copy of the crawled Web to do this at a reduced cost. However, even a search engine cannot scalably construct, store, update, and export a large contextual crawl for an arbitrary set of topics and contexts. This approach is not scalable even for large search engines that have vast resources at their disposal.

The third approach of bootstrapping the crawl with search engine results partially addresses the scalability issues of the second approach, but still has several problems. The first problem is that existing ranking algorithms such as PageRank are global ranking functions across all pages. This means that given several different topics as queries, the same pages with high PageRank reappear as the top results regardless of their rank relative to the target topic. This issue arises twice: first when pages are being crawled, and a second time when pages are presented to the user. For these two cases simple solutions are, respectively, to filter during the crawl and re-rank after indexing the pages. The problem with crawling even given a relevant starting point is that it is not guaranteed that subsequently crawled pages are still relevant. A simple improvement to this approach would be to request more relevant pages from the search engine. While this works to some extent, once context and language constraints are added to the query the number of relevant pages may drop dramatically (e.g. the number of English pages on "malaria" indexed by Google is upward of 13 million pages, but in Serbian there are only 10,200 pages), meaning either some amount of crawling of the web or translation of pages will be necessary. Despite these obstacles this final approach outlines the starting point for our system.

To summarize, the high level design constraints on the solution to the problem are as follows: First, the information stored in the cache should be context appropriate. Second,

the solution should be low cost and scalable; each crawl must be efficient in terms of useful versus useless pages downloaded. Third, the final set of pages in the CIP should not only be searchable but also *browseable* in some fashion.

Research Challenges

Given the design constraints, building a Contextual Information Portal for an arbitrary topic requires us to address several different learning problems in document classification, focused crawling, local search, location sensitivity, machine translation, and copyright detection. We outline each of these problems and discuss the variances that make them distinct from solved problems.

Topic Extraction: The set of topics for which the CIP is to be about may be defined manually based on knowledge of the users' information needs in a particular locale, automatically based on existing logs such as search requests, or something in between. Automatic ontology generation is a well studied problem. It is also reasonable to simply expect a short list of the user information needs along with a description of the context. For this proof-of-concept we assume that a set of topics is pre-defined, and do not consider hierarchies of topics.

Document classification: For a Web page and a specific topic, how do we determine whether the page is related to the topic or not. While there has been a wealth of work on document classification in general (Dumais and Chen 2000; Gao et al. 2003), personalized or context-aware classification is still an area of active research (Zheng, Kong, and Kim 2008; Guan, Zhou, and Guo 2009). The specific variations of the document classification problem relevant to CIPs are: *locality sensitivity* and *copyright detection*. In this work we used only simple document classification techniques as a baseline for further study.

Language compatibility: India alone has 18 official languages and many unofficial languages and local dialects. The African continent has thousands of languages. In contrast, most of the Internet content is still in English, Chinese, and Spanish. Automatically translating pages from one language to another is a large field of research in and of itself, and a key challenge in bringing the Web to the developing world. The distinguishing feature for development is that the corpus of training data for these local dialects is small which presents an additional challenge for statistical methods.

Focused crawling: The goal of a *focused crawler* is to crawl only the relevant portion of the Web that relates to the topic while minimizing the waste of downloading unrelated pages. There is extensive literature on focused crawling related to search engine optimization (Almpanidis, Kotropoulos, and Pitas 2007; Zheng, Kong, and Kim 2008; Pandey and Olston 2008). The main challenge is deciding which of the pages in a crawl frontier are useful to follow.

Information organization: Purely generating a list of pages relating to a topic and indexing them is a complete solution without some level of organization. The goal of organizing the information overlaps with automatic ontology generation. If an ontology is available simply organizing the final set of pages in the CIP based on the ontology may be sufficient. Otherwise, an alternative approach could be to recluster the documents in a more meaningful manner. As an aside, an interesting consideration is whether a high amount of interconnectedness of pages in the CIP is desirable and should be a target of optimization during focused crawling.

Building a Contextual Information Portal

Step 1: Defining topics

There are a multitude of ways to define topics, for this proof of concept we manually picked out a set of topics relevant to the motivating scenarios mentioned at the beginning of the paper. We narrowed down the set of topics to "malaria", "HIV", "calculus", and "organic farming". In some scenarios the user information needs may not fall into large verticals. In those cases, there are a multitude of automated approaches to extract topics from existing request logs. These topics may then be combined with contextual information to derive a broader set of smaller verticals.

Step 2: Training a classifier

After the topics have been defined, the document classifiers must be trained. We initially tried several well known TF-IDF based textual similarity classifiers (Manning, Raghavan, and Schutze 2008), but found that thresholds for these similarity metrics were either too high causing very few acceptable pages or too low causing topical drift. Instead, we opted to train two simple statistical classifiers, a naive Bayesian classifier and a Fisher classifier. To train each classifier we leverage a social bookmarking site *delicious.com* wherein users tag Web pages with relevant keywords. To train each topic we downloaded 5000 pages from *delicious.com* tagged with the name of each topic, and marked these as *good*. We then downloaded 5000 pages from *delicious.com* tagged with "the" and marked these as *bad*. It is also possible to simply train using the top 1000 pages returned by a search engine, but in this work we consider the human tagged pages from *delicious.com* to be the ground truth. We currently ignore contextual cues and simply classify based on whether or not pages fall within a topic.

Step 3: Crawling the Web

To perform the actual crawling process, we implemented a focused crawler based on the Shark crawler (Hersovici et al. 1998). Shark is a heuristic crawler that orders the nodes to be visited by sorting them according to their similarity to their ancestors in a priority queue. At a high level our crawler includes two main modifications to the basic Shark crawler. First, our crawler is bootstrapped with an initial set of A *authoritative documents or authorities* from a large search engine. Second, we add a heuristic technique called *tunneling* (Bergmark, Lagoze, and Sbityakov 2002) that allows a number of irrelevant pages to be traversed to reach a relevant page. Several other modifications are outlined below:

- Input parameters: A, the number of initial authorities, D, a tunneling depth parameter (currently set to 3), and MAX, the quota for the number of relevant documents to retrieve.

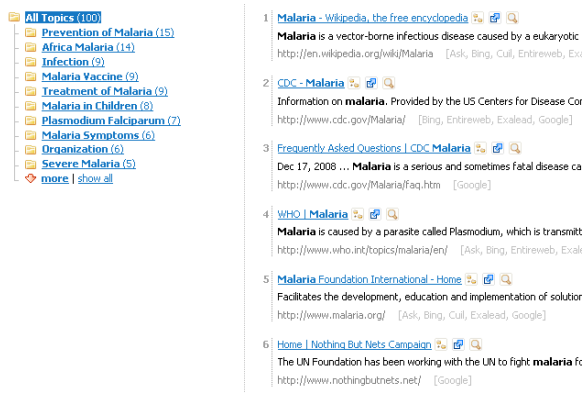

Figure 1: Contextual Information Portal User Interface

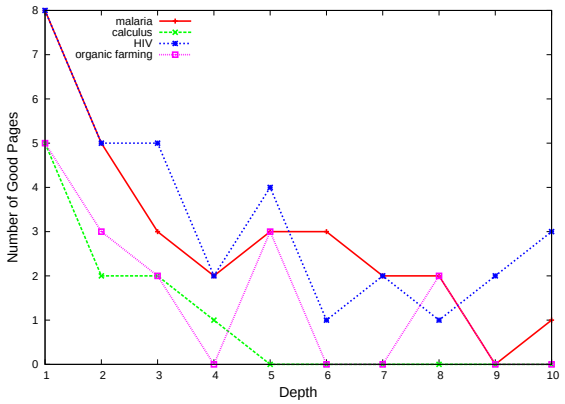

Figure 2: Harvest Rate vs Crawl Depth

- Issue a query to a large search engine to obtain A authorities on topic T. Classify each authority with our classifier for a relevance probability p and insert it into a priority queue q with priority p. Restrict the classification to the top 25% of the feature set normalized by frequency of occurrence.
- The inherited score of a child node, c, of a relevant current node depends on, the probability p assigned by c as the relevance of the current node.
- Anchor text around a hyperlink is obtained by extracting text from a fixed number of surrounding markup elements.

Step 4: Putting it all together

To index, rank, and present the CIP contents we used the free Nutch search engine to index the pages based on textual similarity. We also cluster the search results using the Carrot2 clustering plugin and present these results in a separate user interface pane. A screenshot of the user interface is shown in Figure 1. Users may search the portal and navigate via clicking on links or the topic clusters. All results are cached locally at the CIP and require no network connectivity.

Preliminary Results

As a cursory evaluation of our system, we performed several experiments to assess the scalability of our system and how it performed across a variety of topics. All results here are using our Fisher classifier. We also conducted an informal field study to determine whether our system was actually desired and/or useful.

Experiments

In this simple experiment we wanted to get a sense of how topic drift occurs as a function of their distance from a relevant page. To do this we performed a breadth first crawl from an initial set of relevant pages returned by a search engine. At each iteration we classify the pages in our current set and construct a frontier set by randomly selecting 10 pages linked to by the pages in our current set. We initialize the set of pages with 10 search result pages (depth = 1), and

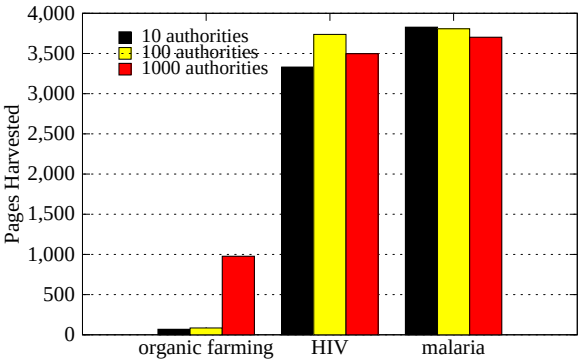

Figure 3: Total Harvested Pages vs Topic & Number of Authorities

ignore duplicates and pages we have already visited. The results are shown in Figure 2. In general the number of good pages found decreases as depth increases, confirming that a simple breadth first search is likely to return many useless pages as the crawl progresses. For topics such as "malaria" and "HIV" for which there are general a lot of large components of connected pages a few pages are found even at depth 10. For topics with fewer relevant pages such as "calculus" no relevant pages are found past a depth 4. This suggests that a more intelligent crawling method is required, particularly when the topic is sparse.

To get a picture of how our focused crawler performs we crawled for pages relevant to "HIV" using 10 authority pages. We found that the harvest rate for our Shark crawler remained relatively stable over the course of the crawl (60% harvest rate) since Shark was able to find relevant pages and crawl in the right direction. In comparison, the BFS harvest rate was nearly 0% after the first three degrees except for a few small successes where relevant pages were discovered by chance. This corroborates the results from Figure 2.

Next, we consider varying the number of authority pages that our crawl begins with. This experiment is to determine whether building a contextual portal is simply a matter of asking a search engine for enough results. Figure 3

shows the final number of relevant pages harvested over 5000 crawled pages for a varying number of initial authorities and topics. In general, the focused crawler is able to retrieve a large number of useful documents regardless of the number of authorities it is initially given, particularly for topics that have a large number of relevant pages overall. However, in the situation where the overall number of relevant pages is few (e.g. "organic farming") more results from the search engine is able to bootstrap the set of pages to be downloaded. This is as expected because the large search engine has many more pages crawled and indexed than our crawler. The results here are a bit skewed by our classifier; for the top 100 authorities our classifier only classified 15 pages returned by Google as relevant for "organic farming" compared to 35 for "HIV" and 53 for "malaria". This suggests that for a given classifier the overall number of useful pages that are classified as "good" may be very limited or not well connected to each other. The implication is that these sparse topics require longer crawls to achieve a particular quota. It would be interesting to experiment with increasing the tunneling depth parameter for these sparse topics to further improve the overall crawl performance.

Initial Feedback

As a first step towards understanding applicability on the ground, we built simple CIPs for two topics (requested by an NGO): *organic farming* and *water harvesting*. We took specific care to remove all copyrighted and non-cacheable web pages from the portal (by using domain specific information and copyright notices in pages). We shipped both portals with 4GB USB sticks as part of a site-visit conducted by an NGO volunteer to agricultural universities and NGOs in India. The system was shown to a experts including: an organic farming scientist, an educational expert, an NGO, students from an agricultural university and some villagers. We briefly describe the positive and negative feedback we obtained from our initial ground experiences.

Positive feedback: An organic farming scientist in India was excited by the idea and mentioned that they definitely need a separate portal for organic farming. Such a portal would help them in better training farmers on organic farming practices as part of a program that they are expanding throughout the country. An educational researcher and the NGO mentioned that the CIP idea would broadly useful to teach pre-school children by gathering all techniques on teaching methods. Teachers could use this database to give project ideas to students. The NGO mentioned that such a portal would help them prepare modules easily on specific sub-topics on demand. An educational expert wondered whether we could build a portal around message postings where users comments from a variety of organizations comment on different but possibly related topics and the portal compiles related postings and responses. Such portals in the context of agriculture are being built on a large-scale as part of aAqua project.

Negative feedback: Based on feedback from individuals in villages (who speak English), we found that our portals would not be useful unless it is in the local language. In addition, the content has to be very location specific for it to be locally useful. One good example in *water harvesting* is a search query for "check dams" - an important water harvesting practice that is being promoted in India. While our portal consisted of several highly ranked pages on organic farming, it performed poorly for "check dams" and did not provide relevant information on how to construct check dams. We learned later that most authoritative pages on water harvesting contained little information on this topic.

The main feedback we obtained from our volunteer who conducted these studies was that:

"Organizations, professionally run NGOs, educational institutions would like this idea but individuals, grassroots organizations working closely with the locals (hence vernacular becomes critical) may not take to this immediately. In both cases though, the relevance factor of the information returned is very important."

Related Work

There is a lot of literature on Web crawling algorithms (Kleinberg 1999; Najork and Wiener 2001; Bharat and Henzinger 1998). Focused crawling was defined and formalized by (Chakrabarti, van den Berg, and Dom 1999), which introduced taxonomic classification or text analysis to determine document relevance and distillation or link analysis to identify authoritative sources of relevant documents. Shark is one of the earlier focused crawlers, and more sophisticated variations exist (Peng, Zhang, and Zuo 2008; Pandey and Olston 2008). Also, extensive follow up work has compared focused crawling against a variety of other crawling techniques across a variety of domains (Menczer et al. 2001; Davison 2000; Cho, Garcia-Molina, and Page 1998). It has been shown in recent work (Almpanidis, Kotropoulos, and Pitas 2007) that uses a latent semantic indexing (LSI) classifier, which combines link analysis with text content, that a variant of the simple Shark-search can be surprisingly efficient when compared to more sophisticated and expensive techniques such as LSI and PageRank (PR).

A wide variety of document classifiers have been developed: Decision Tree (Gao et al. 2003), Support Vector Machines (Dumais and Chen 2000), taxonomic classifiers (Chakrabarti, van den Berg, and Dom 1999), and many others (Zheng, Kong, and Kim 2008). These document classifiers typically depend on positive and negative examples for training. While taxonomic classifiers have the benefits of being train-once, use-almost-everywhere generality and also higher accuracy in many cases we opted to explore simpler classifiers for two reasons which appear to mirror each other. First, the comparatively inexpensive training cost taxonomic classifiers. Second, once context constraints are included, the scarcity of positive training data will become a serious concern for some of these classifiers (Gao et al. 2003). However, recent text advances in centroid-based classifiers appears to satisfy both of our constraints while maintaining good performance (Guan, Zhou, and Guo 2009).

The application of focused crawlers to particular topics in an effort to build digital libraries and web caches has been considered before (Bergmark, Lagoze, and Sbityakov 2002; Ehrig and Maedche 2003; Qin, Zhou, and Chau 2004).

The underlying crawling and classification technology behind building digital libraries is mature, but unique challenges and considerations arise in the context of development (Isaacman and Martonosi).

Conclusions and Future Work

In this work we motivated and defined the concept of a Contextual Information Portal and considered the unique requirements for its construction. To understand the limiting factors and design space we implemented a proof-of-concept. We demonstrated how a Contextual Information Portal may be constructed with very limited computing resources. We found that for a portal in English, simple off-the-shelf solutions were sufficient. However, once constraints such as a specific location or language are included, the set of candidate pages falls dramatically.

From our initial feedback we hope to incorporate context specific information to enhance the appropriateness of the pages in our portal. As an immediate next step we will investigate how improvements to our classifier and the inclusion of contextual information in our training data will affect the quality of our results. Our problem touches on many other areas related to artificial intelligence including: topic extraction, crawling, machine translation, and document clustering algorithms. Finally, it would be interesting to explore cost-efficient methods of updating the portal via low bandwidth links or large updates over longer timescales by physically shipping updates via high capacity storage media.

References

Almpanidis, G.; Kotropoulos, C.; and Pitas, I. 2007. Combining text and link analysis for focused crawling-an application for vertical search engines. *Information Systems*.

Bergmark, D.; Lagoze, C.; and Sbityakov, A. 2002. Focused crawls, tunneling, and digital libraries. *Lecture notes in computer science*.

Bharat, K., and Henzinger, M. R. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st annual international ACM SIGIR conference*.

World health organization - blue trunk libraries. http://www.who.int/ghl/mobile_libraries/bluetrunk/en.

Brewer, E.; Demmer, M.; Ho, M.; Honicky, R.; Pal, J.; Plauche, M.; and Surana, S. 2006. The challenges of technology research for developing regions. *Pervasive Computing, IEEE*.

Chakrabarti, S.; van den Berg, M.; and Dom, B. 1999. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networking* 1623–1640.

Chen, J.; Subramanian, L.; and Li, J. 2009. RuralCafe: web search in the rural developing world. In *Proceedings of the 18th international conference on World wide web*.

Cho, J.; Garcia-Molina, H.; and Page, L. 1998. Efficient crawling through url ordering. In *Proceedings of the seventh international conference on World Wide Web*.

Davison, B. D. 2000. Topical locality in the web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*.

Dumais, S., and Chen, H. 2000. Hierarchical classification of Web content. In *Proceedings of the 23rd annual international ACM SIGIR*, 263.

echoupal. http://www.itcportal.com/rural-development/echoupal.htm.

Ehrig, M., and Maedche, A. 2003. Ontology-focused crawling of Web documents. In *Proceedings of the 2003 ACM symposium on Applied computing*.

Gao, S.; Wu, W.; Lee, C.; and Chua, T. 2003. A maximal figure-of-merit learning approach to text categorization. In *Proceedings of the 26th annual international ACM SIGIR*.

Guan, H.; Zhou, J.; and Guo, M. 2009. A class-feature-centroid classifier for text categorization. In *Proceedings of the 18th international conference on World wide web*.

Hersovici, M.; Jacovi, M.; Maarek, Y. S.; Pelleg, D.; Shtalhaim, M.; and Ur, S. 1998. The shark-search algorithm. an application: tailored web site mapping. In *Proceedings of the seventh international conference on World Wide Web*.

Isaacman, S., and Martonosi, M. The C-LINK System for Collaborative Web Usage: A Real-World Deployment in Rural Nicaragua.

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

Manning, C. D.; Raghavan, P.; and Schutze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Menczer, F.; Pant, G.; Srinivasan, P.; and Ruiz, M. E. 2001. Evaluating topic-driven web crawlers. In *Proceedings of the 24th annual international ACM SIGIR conference*.

Mubaraq, S.; Hwang, J.; Filippini, D.; Moazzami, R.; Subramanian, L.; and Du, T. 2005. Economic analysis of networking technologies for rural developing regions. *Workshop on Internet Economics*.

Najork, M., and Wiener, J. 2001. Breadth-first crawling yields high-quality pages. In *Proceedings of the 10th international conference on World Wide Web*, 114–118. ACM New York, NY, USA.

Mit opencourseware. http://ocw.mit.edu.

Pandey, S., and Olston, C. 2008. Crawl ordering by search impact. In *Proceedings of the international conference on Web search and web data mining*, 3–14. ACM.

Peng, T.; Zhang, C.; and Zuo, W. 2008. Tunneling enhanced by web page content block partition for focused crawling: Research Articles. *Concurrency and Computation: Practice & Experience*.

Qin, J.; Zhou, Y.; and Chau, M. 2004. Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*.

Web initiative for surgical education. http://wise-md.med.nyu.edu.

Zheng, H.; Kong, B.; and Kim, H. 2008. Learnable focused crawling based on ontology. *Lecture Notes in Computer Science*.