Université de Montréal

Domain Specific Web Search

par

Zheng Zhen

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Maîtrise ès sciences (M.Sc.)
en informatique

Décembre 2004

Université de Montréal

Faculté des Études Supérieures

Ce mémoire intitulé :

Domain Specific Web Search

présenté par:

Zheng Zhen

a été évalué par un jury composé des personnes suivantes :

Guy Lapalme

Président Rapporteur

Jean Vaucher

Directeur de recherche

Jian Yun Nie

Membre du Jury

Mémoire accepté : 11 janvier 2005

Résumé

Un nouveau modèle de recherche d'information développé dans ce mémoire traite des problèmes de précision et complétude pour les systèmes de recherche courants. De nos jours, la recherche d'information sur le Web est devenue de plus en plus importante dans toutes les applications informatiques. Il y a beaucoup de moteurs de recherche commerciaux puissants tels que Google, Askjeeve, Yahoo et de MSN etc… Tous ces systèmes de recherche semblent pouvoir accepter des requêtes arbitraires, avec un secteur de recherche large. Cependant, leurs résultats manquent de détails dans la plupart des situations. Ce mémoire se concentre sur une situation différente: des recherches récurrentes dans quelques domaines ciblés. Ceci exige des résultats de recherche plus précis et plus complets.

Pour créer un tel système, nous avons utilisé plusieurs progiciels du domaine public et nous présentons un survol des systèmes disponibles.

Comme expérimentation finale, nous avons utilisé notre application de recherche pour trouver des appels d'offres de 5 villes canadiennes choisies aléatoirement. Les résultats prouvent que nos techniques fonctionnent mieux que les moteurs de recherche conventionnels.

Mot-clé :

domaine spécifique, recherche, recherche documentaire

Abstract

A new information-searching model is developed in this thesis to deal with the precision and completeness problems for current searching systems. Nowadays, web searching has become more and more important in all computer applications. There are many powerful commercial search portals such as Google, Askjeeve, Yahoo and MSN; these searching systems support arbitrary queries, with a wide search area, and their results lack in depth in most situations. This thesis focuses on a different situation: recurrent searches in some specified domain. This requires the search results to be more focused.

To create such a system, we utilized several public domain packages and we present a survey of available systems.

As a final experiment, we tested our search application to find call for tenders (CFT) from 5 randomly selected Canadian cities. Results show that our techniques work much better than conventional search engines.

Keyword:

domain specific, search, information retrieval

List of Tables

List of Figures

Acknowledgement

Special thanks are due to my supervisor, professor Jean Vaucher. His help and advice enabled me to improve and complete this work. With his knowledge and straightforward way he has inspired me to do my best.

I am also gratful to all the professors and staff members in the department of computer science and operations research, for their efficiency, professionalism and sympathy.

This work has been carried out with major support from NSERC as a joint research project with CIRANO and Nstein Technologies. I have benefitted from many discussions with Stéphane Vaucher and other MBOI team members. I thank all concerned.

Finally, very special thanks goes to my wife for her support and encouragement in making this dream coming true.

Chapter 1
Introduction

For the procurement of expensive goods and services, governments and corporations rely on the call-for-tender process: publishing descriptions and conditions for what they need and selecting from submitted tenders. Currently, the trend is to publish Call For Tenders (CFTs) on the Web and many government bodies have set-up major tendering sites which list thousands of CFTs from their various departments. In Canada, MERX (http://www.merx.com) is an example; in the US, FedBizOpps.gov is the single government point-of-entry (GPE) for Federal government procurement opportunities over $25,000; TED (Tenders Electronically Daily) is the equivalent for the European community. Other organizations, such as SourceCan (http://www.sourcecan.com), provide an aggregation service by collecting and indexing documents from other sites.

A quick look an SourceCan (October 2004) shows clearly the importance of this phenomenon: it shows 3400 open bids from Canadian sources, 14,000 from US governments sites, 33,000 European Union public tenders, 2,100 for the World Bank and over 20,000 more opportunities from other sources. However, this abundance of listings hides very real problems when it comes to exploit the data. First, there is no unique standard format for these documents. Most are loosely structured and meant for human interpretation; some documents are quite detailed whereas others are limited to a one-line title. Secondly, product codes – when they are used – may follow different standards. Indexing is therefore partial and approximates. The sheer volume means that effective monitoring of these sites must be done automatically; but the automatic classification of CFTs and identification of pertinent opportunities is difficult.

This situation motivated the launch of a joint university / industry research initiative with researchers from the Universite de Montreal, Nstein and CIRANO. The project, “Matching Business Opportunities on the Internet”, is aimed at developing systems which can find and classify CFTs accurately. It has several facets: one deals with the design of collector agents that systematically visit known sites and download new documents. Another, the biggest, concentrates on the natural language analysis and classification of documents [F. Paradis et al. 2004]. The part which is the topic of this thesis is the design of discovery agents responsible for finding and mapping new CFT sites so that they can be exploited by the collector agents.

In spite of the volume of data concentrated in the major sites listed above, it is important to continuously scour the Web for others. One reason is that many CFTs linked to local governments, large companies, utilities (like Hydro-Quebec) or particular industrial sectors (like Construction, Transportation or Wood products) are published on local sites and not listed on the major ones. This was confirmed by our initial studies. During the last year and a half, the MBOI collector robots visited about 35 sites daily in addition to the major sites. These are especially important because for many companies, CFTs closer to home are easier to take advantage of. A final justification for efficient continuous discovery is that sites are often reorganized and the scripts used by the collector agents no longer work. Over a year, we found that about half the sites changed and we had to rediscover how to extract CFTs from them.

Our discovery agents are basically web spiders, but they target specific kinds of documents and the techniques they require are quite different from those used for commercial search engines (CSEs) like Google or Yahoo. Those systems cater to a heterogeneous clientele and their coverage of the Web is broad but shallow. They need massively parallel search to handle the estimated million pages added to the web every day and must use techniques such as Page Rank [Page et al. 1998] to rate Web pages objectively.

In contrast, we are interested in targeted search with a narrow but deep focus. In particular, we would like to search the invisible web accessible only by complex web navigation involving entering data into forms and which is often not indexed by CSEs. Another difference is that our agents are not primarily interested in documents; they are interested in sites which contain documents. This is because the sites are dynamic and the actual job of document retrieval is left to collector agents. A collector agent doesn't need a list of documents to retrieve; rather it needs a script detailing how to navigate a site and identify new CFTs.

Early in the project, it became evident that intelligent discovery entailed the combination of a multitude of techniques that would have to be adapted to the reality of the web: what do CFT pages look like? How are CFT sites organized? What data formats are used? Therefore it was decided to develop a prototype to evaluate the aptness of an initial architecture based on a combination of published techniques. Only once each module achieved a certain level of performance could we start more systematic adaptation and evaluation. The prototype was tested on the task of finding the CFTs of a city government given only the city name as starting point.

This thesis presents what we learned from the prototype. It is organized as follows. In Chapter 2, we review some general principles and processes of information retrieval. In Chapter 3, we briefly review current research and list the techniques that should be incorporated in a focused spider. In Chapter 4 and Chapter 5, we describe the architecture of our system and comment on the public domain software tools that were used in its implementation. In Chapter 6, we present some experimental results that show the performance of the system and led to the heuristics used in various modules. Finally, in Chapter 7 gives some conclusions.

Chapter 2
Information Retrieval

With the explosion of new sciences and technologies, the problem of information storage and retrieval has been receiving increasing attention. To face the vast amount of information, an accurate and speedy access to information is eagerly awaited. However, difficulties come from either missing relevant information or duplicating useless information. With the rise of computers, there have been a great number of ideas about how to provide a rapid and intelligent retrieval system by computers. In this chapter, we look at information retrieval in classical context. Retrieval from the web is considered in the next chapter.

2.1 Information retrieval and Data Base

Database systems are designed to provide efficiently access to large quantity of information. They are not suited for information retrieval (IR) needs.

If we have a database-generated site, such as an e-commerce catalog or a stock exchange for on-line quotations, we might wonder why we need another type of retrieval -- IR. There are some obvious reasons for pursuing a new type of retrieval:

1) Database information is stored in separate fields, but searchers do not want to do searching by choosing fields one by one manually.

2) The response time of database searching can be extremely long, especially for the multiple words searching.

3) The results coming from the database searching cannot be sorted by relevance automatically.

Then, what is so-called IR?

A straightforward definition is given by Lancaster,F.W. [Lancaster,F.W 1968]: “An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request."

In IR, we are more interested in finding those items that partially match the requests, and then we will select the best matching ones. The inference and model used in IR are inductive inference and probabilistic modeling. The preferred query language of IR is natural language and the query may be incomplete and vague. Moreover, in IR we search for relevant documents that sometimes don't have exactly matching items. Some small errors in matching sometimes do not affect the performance of the system significantly.

We will discuss and explain how to achieve the above IR features in chapter 5.We'll also introduce an IR system construction tool -- Lucene there. Lucene is a tool that provides some convincing evidence of its efficiency and effectiveness of those IR features [Apache Jakarta Project 2003].

2.2 Information retrieval process

The typical information retrieval (IR) system is illustrated in figure 2.1. Just like a library's retrieval system, this system includes a storage of documents and books, an index of this huge storage, and a database of the indexed documents. When users, such as a student or a teacher, require some information, they first need to formulate questions, and then send a query to this system. After receiving a query and matching each items stored in the database to this query, the system retrieves the information and replies to the user. When this system is on-line, it is possible for the user to change his requests frequently for different searching purposes. The whole procedure may be repeated and refined for several times, so as to improve the subsequent retrieval results. This repeated procedure is called relevance feedback. From this procedure, we can see that the basic elements of information retrieval are: documents storage, indexing, index storage, query, matching (searching) and feedback.

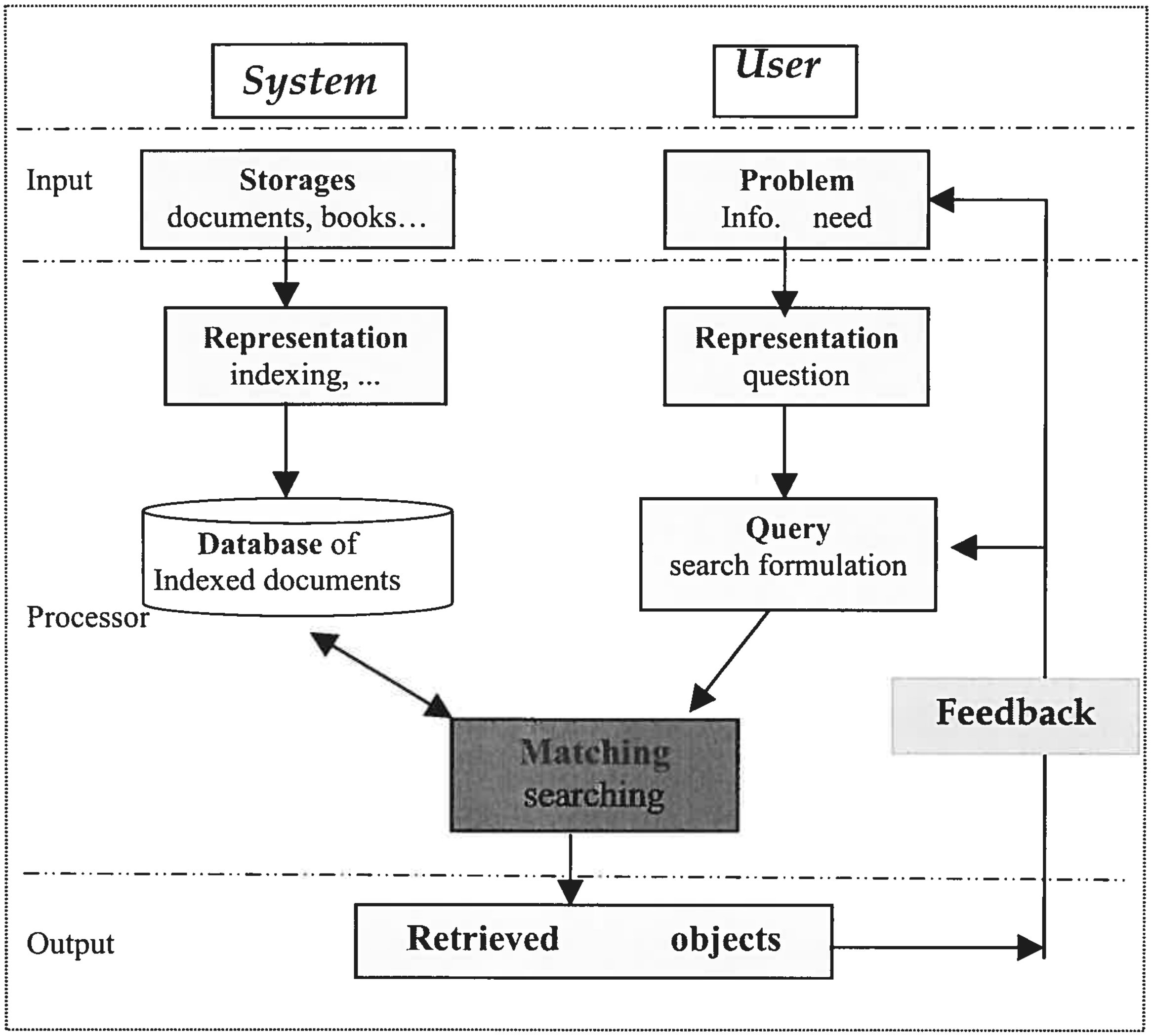


Figure 2.1 *Information retrieval (IR) system*

We can abstract the system in a further step, circling three components: input, processor and output.

First, to start with the input, it can link the document storage and users' problems. It should show the best link or the best relationship between the document storage and users' problems, but it usually does not.

Then second, the processor is the core part of the retrieval system for the retrieval processes. The main problem here is how to obtain representations of documents and queries respectively. A lot of computer-based library retrieval systems don't store original documents or queries, they only store some representations of them. It means that the text of a document is lost, once it has been processed for the purpose of generating its representation. The process may involve structuring the information in some appropriate ways, such as classifying the information. It will also involve performing the actual retrieval function, that is, executing the search strategy in response to a query. In the above diagram, the representation of each document and query has been placed in processor side to emphasize such a fact -- to form these representations is actually an important task during the retrieval process. So the structure of documents is often considered as a part of this process. Finally, the output will come out with a set of references or document numbers.

2.2.1 Indexing process

Indexer transfers those documents into one index file which provides for the next searching.

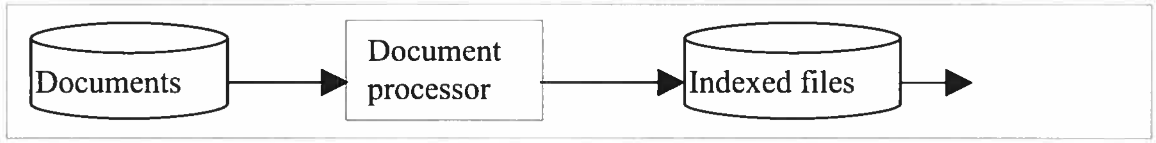


Figure 2.2 *Indexer*

The core of indexer is the document processor, which normally includes 9 steps:

❶ Normalize the document stream to a predefined format.

❷ Break the document stream into desired retrievable units.

❸ Isolate and “metatag” subdocument pieces.

❹ Identify some potential elements in documents for indexing.

❺ Delete “stop words” like: a, an, the, on, at etc.

❻ Stem from terms.

❼ Extract the index entries.

❽ Compute the weights.

❾ Create and update the main inverted file against which the search engine searches to match queries to documents.

Steps ❶,❷, and ❸ are Preprocessing. These first three steps simply standardize the multiple formats encountered when deriving documents from various providers. These steps serve to merge all the data into a single consistent data structure that all the downstream processes can be handled. Step two is important because the pointers stored in the inverted file will enable a system to retrieve units of various sizes — site, page, document, section, paragraph, or sentence.

The fourth Step identifies elements to index. The identification of the potential indexable elements in documents will seriously affect the nature and the quality of the document representation, which the search engine will depend on. To design the system, we must define the word "term". It usually includes the alpha-numeric characters between blank spaces and punctuations. In most way, it should also consider the non-compositional phrases (for example “hot dog"), the multi-word names, and inter-word symbols, such as hyphens and apostrophes, for example, "small business men" versus “small-business men”. Each search engine has a series of rules that are used to define a term suitable for indexing. The document processor will tokenize the input streams according to these rules.

Next Step is to delete “stop words”. This step mainly contributes to save system resources by eliminating “stop words” from the above process. It may comprise up to 40 percent of text words in a document. A “stop word” list typically contains those words known to convey no useful meanings, such as Articles (*a, the*), Conjunctions (*and, but*), Interjections (*oh, but*), Prepositions (*in, over*), Pronouns (*he, it*), and formats of the "be" verb (*is, are*). To delete stop words, an algorithm compares index term candidates in the documents with a “stop word” list, and eliminates them from the documents for searching.

The sixth step is term stemming. Stemming is a process of removing word suffixes, which achieve two goals: 1). To improve efficiency, stemming reduces a number of unique words in the index, reduces the storage space required for the index, and speeds up the search process. 2). For effectiveness, stemming improves recall by reducing all forms of the word to a basic or stemmed form. For example, if users ask for *analyze*, they probably also want documents which contain *analysis*, *analyzing*, *analyzer*, *analyzes*, and *analyzed*. Therefore, the document processor stems document terms to *analy-* so that documents that include various forms of *analy-* will have equal possibilities to be retrieved. Of course, this stemming may negatively affect precision: when users query an exact word format, all forms of a stem will be matched, instead of the formatted word in the query form.

Systems may implement either a strong stemming algorithm or a weak stemming algorithm. A strong stemming algorithm will strip off both inflectional suffixes (*-s, -es, -ed*) and derivational suffixes (*-able, -aciousness, -ability*), while a weak stemming algorithm will strip off only the inflectional suffixes (*-s, -es, -ed*).

The seventh Step is to extract the index entries. Having completed steps ❶ through ❻, the document processor extracts the remaining entries from the original document.

The output of step ❼ is to insert and store in an inverted file that lists the index entries and an indication of their positions and frequencies of occurrences. The specific nature of the index entries, however, will vary based on the decision in Step ❹ , which will concern what constitutes an "indexable term." Some document processors will even have phrase recognizers, for example, Named Entity recognizers and Categorizers.

The eighth Step is Term weight assignment. Weights are assigned to each term in the index file. The simplest search engines only assign a binary weight: 1 for presence and 0 for absence. The more sophisticated the search engine is, the more complex the weighting scheme is. To measure the frequency of occurrence of a term in the document creates more sophisticated weighting. Generally, "tf/idf." is an optimal weighting in IR. This algorithm measures the frequency of occurrence of each term within a document. Then it compares that frequency in one document with the frequency of occurrence in the entire database.

Not all terms are good as "discriminators" — which means, not all terms can single out one document from another very well. A simple example is the word "the". This word appears in too many documents to help distinguish one from another. A less obvious example would be the word "antibiotic". For a sport database, when we compare each document to the whole database, the term "antibiotic" could be a good discriminator, and therefore, it should be assigned at a high weight. Conversely, in a health or medicine database, "antibiotic" could be a poor discriminator because it occurs very common. Practically, the TF/IDF weighting scheme assigns higher weights to those terms that really distinguish one document from the others.

The last step is to create index. The index or inverted file is the internal data structure that stores the index information and will be searched by each query. Inverted files can be a simple listing of alpha-numeric sequence in a group of documents/pages indexed, with an overall identification numbers of the documents in which the sequence occurs to a complex list of entries, the tf/idf weights, and pointers. The more complete the information in the index is, the better the search results are.

We'll give a concrete example in chapter 4, in which Lucene will be introduced as a very good indexing tool for constructing indexer and search engine.

2.2.2 Query process

Query processing has six steps at different level. It shares many steps with document processing. The whole processing is shown in the figure 2.3 below.

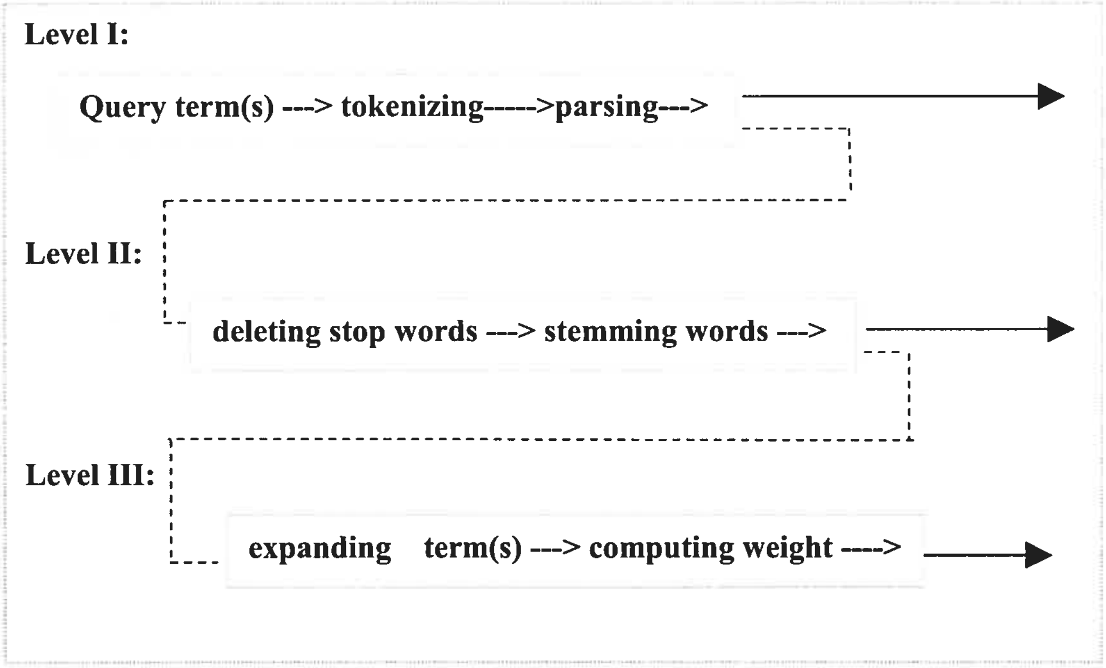

Level I:
Query term(s) ---> tokenizing----->parsing--->
Level II:
deleting stop words ---> stemming words --->
Level III:
expanding term(s) ---> computing weight --->

Figure 2.3 *Query Processor*

Level I:

Tokenizing: As soon as a user inputs a query, the search engine will tokenize the query stream, i.e., break it down into understandable segments. Usually a token is defined as an alpha-numeric string that occurs between white spaces and/or punctuations.

Parsing: Because users may utilize some special operators in their query, including Boolean, adjacency, or proximity operators, the system needs first to parse the query into query terms and operators. These operators may occur in the form of reserved punctuations (e.g., quotation marks) or reserved terms in specialized formats (e.g., AND, OR). At this point, a search engine may take the list of query terms and search them with the inverted file. In fact, this is the point at which most of the available public search engines perform the search.

Level II:

Stop list and stemming: Some search engines will go the stop-list and stem the query, similar to the processes described above in the Document Processor section. The stop list might also contain those words from frequently occurring query/phrases. However, because most of the available public search engines encourage short queries, the engines may drop these two steps. Some particular search engines can create a query representation combining with its matching function and perform the search against the inverted file at this point.

Level III:

Query expansion: Because most of the users usually search some single statement of information with a query, it is highly probable that we can use synonyms as a substitute, not be limited on the exact query terms, to search in the documents for meet users' needs better. Therefore, more sophisticated systems may develop the query into all possible synonymous terms --- even broader and narrower terms.

Query term weighting (assuming more than one query term): The final step in query processing is to compute weights for each term in the query. Sometimes those users control this step by simply indicating either pre-fixed/required each term's weights or which term in the query must be considered in each retrieved document to ensure relevance. After this final step, the expanded, weighted query is searched against the inverted file of documents.

Although a system can choose any one of the 3 kinds of query processors to match the query to the inverted file, numerous steps and documents make this process more expensive for processing with the computational resources and responsiveness. However, the longer the waiting time is, the higher the quality of results will be. Thus, there is a tradeoff between time and quality. Available public search engines usually choose speed, not quality, having too many documents for searching. Sometimes for some special case, users may occasionally obtain better quality, for example, in domain specific web search.

2.3 Areas of search

Effort to improve information retrieval (IR) can be subdivided in content analysis, information structure and evaluation. Briefly, the content analysis is concerned with describing the contents of documents in a form suitable for computer processing. The information structure is concerned with exploiting relationships between documents to improve the efficiency and effectiveness of retrieval strategies, and the third one -- evaluation is concerned with the measurement of the effectiveness of the retrieval.

Usually we use the vocabulary-usage frequency in the document text to determine which word is sufficiently significant to represent or characterize this document in the computer. Thus, a list of what might be called 'keywords' derive form each document. In addition, the frequency of the occurrence of these words in the text body could also be used to indicate a degree of significance. This provides a simple weighting scheme for the 'keywords' in each list and makes a representation in the form of a 'keyword description'. The use of statistical skill about distributions of words in documents reveals statistical associations between keywords. These associations can provide a base for the construction of a thesaurus – an aid to information retrieval (IR). The frequency that any two keywords occur together in the same document is called co-occurrence frequency. It can be used to measure of the association between keywords [Baroni et al. 2002].

Specifically, the information structure is a logical organization of information. The development of information structure should empasize on the logical structure of the information (documents, files). Some earlier experiments for those document retrieval systems usually adopted a series of file organizations. Up to now, it is more often to use the inverted file or clustered files for on-line retrieval.

Evaluation of retrieval systems historically has been proved extremely difficult. As early as 1966, Cleverdon listed six main measurable quantities[Cleverdon et al. 1966]:

(1) The *coverage* of the collection, that is, the extent of the system included relevant matters;

(2) The *time la*g, that is, the average interval between the spending time of the search request and finding out an answer;

(3) The form of output *presentation*;

(4) The *effort* involved on the part of the user in obtaining answers to his search requests;

(5) The *recall* of the system, that is, the proportion of relevant material actually retrieved in answer to a search request;

(6) The *precision* of the system, that is, the proportion of retrieved material that is actually relevant.

There are two factors: the recall and the precision that are used to measure the effectiveness of the retrieval system. In other words, they are the rulers of the system abilities to retrieve relevant documents and to hold back the non-relevant ones at the same time. The more effective the system is, the more it will satisfy the user. It is assumed that the precision and the recall are sufficient for the measurement of effectiveness.

Nowadays, all the IR sub-categories --- content analysis, information structures and evaluation, have been already applied in computerized IR systems, in on-line IR systems, and even in the web searching systems. For example, we can give a keyword to calculate those documents' or web pages' tf / idf, so as to determine the most relevant files or pages. In order to make the information structure for those documents or web pages, computer should get the logical structure of the documents/web pages first [Carchiolo et al. 2003].

Although, web pages have a lot of differences with traditional documents, the basic principles of IR can also apply in the web. Chapter 3 will mainly discuss a situation of web search --- IR on the web.

Chapter 3
Survey of Web Search

The web search problem actually is a problem of IR on the web. The whole web search process basically includes three parts: search engine, indexer and crawler [Sherman et al. 2003]. Different users may pay attention to different parts. One can complete its own page-ranking algorithm, while others may add more function to their crawler parts [Brin et al. 1998]. Referring to the figure 3.1, following the order from top to bottom, we'll go into details for each part of web search one by one.

Generally, as an important concept, a crawler-based searching system has three basic elements [Crimmins 2002]. One is that the crawler downloads the new web pages from the web continuously. The second is that the indexer gathers the words from those documents, whether they are HTML pages, local files or database records, and the indexer picks up those words into one index for a fast information retrieval. The third element is the search engine itself, which accepts the queries, locates the relevant pages in the index and formats the results in an HTML page (see figure 3.1). Indexer and Search engines are heavy-duty server programs, thus it is required that processors be fast for the processing, hard-disk space be significant for the index and a great deal of bandwidth must be available for responding to many simultaneous searching requests. An exact configuration depends on the numbers of pages, but most of the search engines require Intel Pentium or Sun Solaris processors, Microsoft Windows NT/2000 or Unix, and at least a T1 line.

For those specific domain search applications [Oyama et al. 2004], an optimal solution can make a crawler more powerful and smarter by making it able to follow extended links intelligently while it is crawling on the fly.

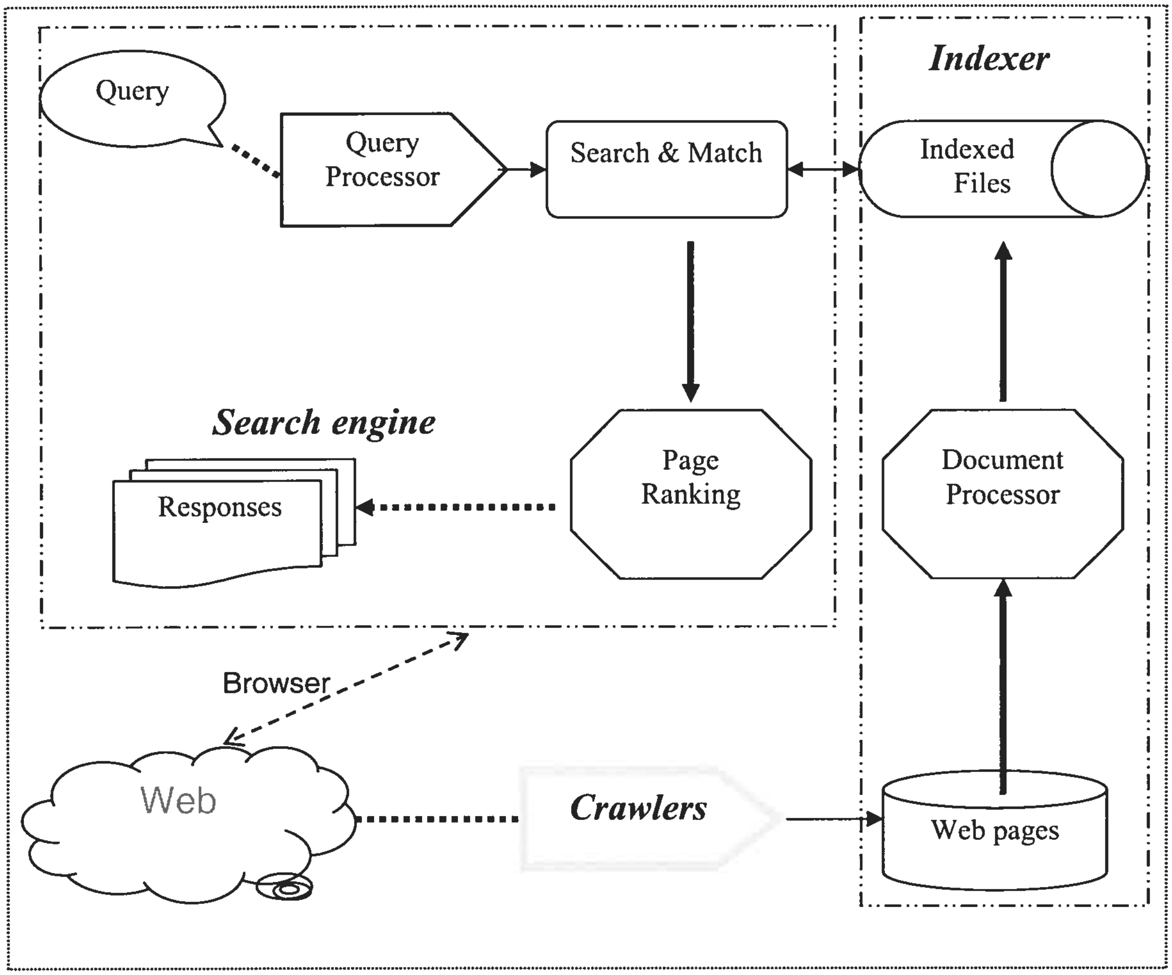


Figure 3.1 W*eb search system*

The term "specific domain", it does not only mean a domain specified before searching, but also mean a domain specified with certain arbitrary. For a domain specific searching, we can easily apply heuristic search. Thus, the crawler is able to crawl with a focused crawling to reduce the search space and the storage of downloaded pages and indexed files. With the help of the basic crawler-based search engine technologies, we develop a domain specific information searching system on the base of the huge web resources.

3.1 Crawler

A web crawler is a program which automatically traverses the web by downloading documents and following links from page to page [Koster 1999]. In the input side, the features of the downloaded web pages can highly influence the behavior of crawler. The workflow of crawler looks like the following chart.

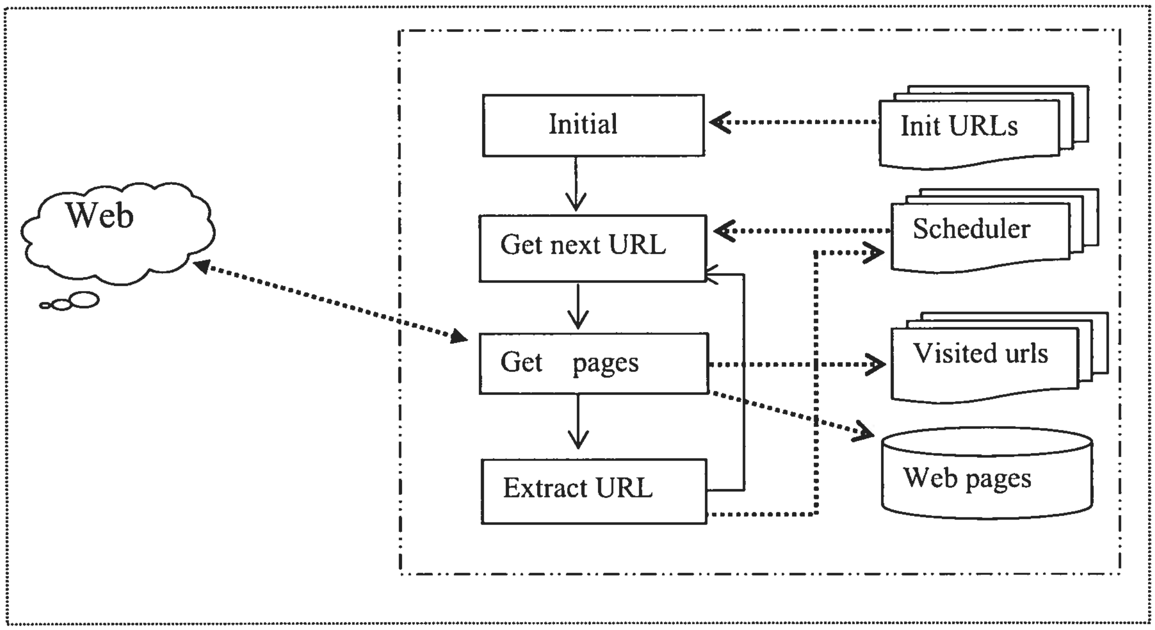


Figure 3.2 *Crawler*

After the crawler gets an initial URL, it will visit this URL and download this web page. It extracts URLs from downloaded pages, sends extracted URLs to the scheduler, and continuously begins the next download loop. The scheduler will be responsible for choosing a URL from those extracted URLs for guiding the crawler on the web in its further step. This is the basic continuous process. In reality, we often complement its functions with intelligent crawling.

Among all the web search systems, some crawlers are indiscriminate crawlers, which means that they will download web pages without any choosing, and any new pages are possible to be downloaded. But there are still some others that can follow some rules when crawling [Mladenic 1999]. Some interesting methods proposed are those of *Fish Search* [Bra 1994] and *focused crawling* [Chakrabarti et al. 1999]. The essential idea in focused crawling is that there is a short range topical locality on the web. This locality may be used in order to design effective techniques for resource discovery by starting at a few well chosen points and maintaining the crawler within the ranges of these known topics. These crawlers only download new pages with an interesting topic. Hub pages (web pages containing links to a large number of pages on the same topic) and authority pages (documents whose content corresponds to a given topic) may be used for the purpose of such crawling. Generally, the hubs on a subject are likely to point to authorities and vice-versa. In addition, structural properties of world wide web like linkage locality (web pages on a given topic are more likely to link to those on the same topic) and sibling locality (if a web page points to certain web pages on a given topic then it is likely to point to other pages on the same topic) can be applied in crawling [Kleinberg 1998].

Besides, to implement a focus crawling, a crawler must create some learning modes, in other words, it should have some intelligence [Aggarwal et al. 2001]. There are 6 learning modes proposed by Aggarwal:

1) Probabilistic mode for priorities
2) Content based learning
3) URL token based learning
4) Link based learning
5) Sibling based learning
6) Combining the preferences

Dr. J.H. Cho also gave some more complex methods through URL analysis in 1998[Cho et al. 1998]. The advantage of URL analysis is that it is fast and light-weight (don't rely on much storage) so that a crawler can apply it on the fly.

Since some learning modes are based on much calculation, only part of them are proper for creating a light-weight crawler, such as mode 1), 3), 4) and mode 5).

When wandering on the web, crawler can explore new URL with different algorithm: breadth-first, depth-first or best-first. Whatever breadth-first or depth-first, it must go through all the web pages (nodes) to finish a complete solution. However, it is obviously impossible to crawl all the 500 billions pages on the web for each searching, even for Google, which can only crawl and store 5 billions web pages. To try a heuristic function, the best-first search algorithm is an optimal solution. If a crawler can keep users' keyword in mind, explore new web pages along a route close to the users' interest, it will come out with what we want. Web searching means a focusing crawling that is able to crawl specific topics on portions of the web quickly without having to explore all web pages. To realize this selective crawling, a crawler should be able to select the next node based on some criteria.

3.2 Search engine

Search engine working process looks like the following chart. It mainly includes query processing, search and match process, and page ranking[Elizabeth Liddy 2001].

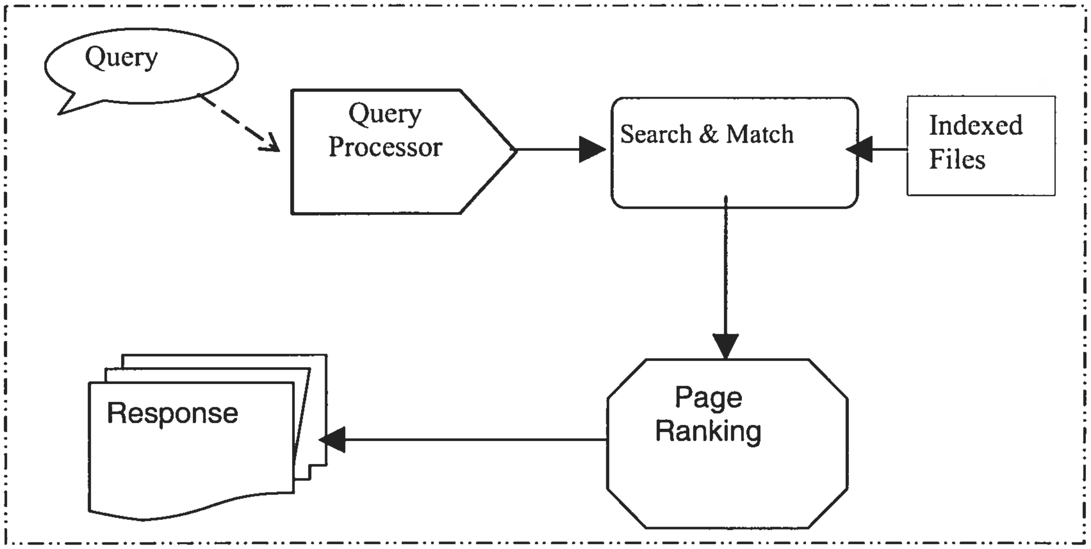


Figure 3.3 *Search engine*

3.2.1 Search and match function

Search and match function depends on the information retrieval mode of the searching system. About the searching the inverted file for documents to meet the query requirements, please refer to the above "matching" process. While the required computational processing for simple, un-weighted, non-Boolean query matching is far simpler than a weighted, Boolean model. The search and match function is related to each other clearly and easily, as the search system information retrieval model, the query processor model and the document processor model are determined above.

3.2.2 Page ranking

To determine the degree to which those subsets of documents or pages match the query requirements well or not, a similarity score can be computed between the query and each document / page, based on the scoring algorithm used by the system. Scoring algorithms rankings are based on the presence of query terms, term frequency, tf/idf, Boolean logic, and query term weights. Some search engines use scoring algorithms not only based on document contents, but also based on the relationship among documents or past retrieval history of documents /pages. After computing the similarity of each document in the subset of documents, the system provides an arranged list to the user. Some systems' orders also depend on query weighting mechanisms. However, after the search engine determines ranks, the user can simply click and follow hyperlink to the selected document /page. More sophisticated systems, as figure 2.1 shows, will allow the user to provide some relevance feedback or to modify the query based on the results received.

3.3 Performances of a web search system

General speaking, web search has these basic characters:

1) web search has to face heterogeneous storage with various format, language etc;

2) web pages are never indexed on the web;

3) web search face an open, dynamic, and variable situation. The content of a specific web page may be changed periodically, or even totally disappear.

4) all the pages on the web are connected only by hyperlink within pages. They are absolutely in a huge network.

In order to discuss the web search, we modify the Figure 3.1 to Figure 3.4. To improve performances of a web search system, a crawler of a search system can extend its function for different designs. We will discuss this more in the following chapters. The processor is a pure IR application, basically the same as the traditional IR system. For the output layer, there is an additional page ranking procedure. For any web search, matching is partial match or best match; given a proper query, it is quite easy to find numerous responses to meet users' requests on the web; any web search is an open, dynamic process, i.e., users can browse webs following the search results, and any bias or error may lead users to browse outside of his concentration. Therefore, an important step of the page ranking is to ensure that users get partial but best match corresponding his/her query. This is also an example of evaluation – the most difficult one of the three IR subcategories.

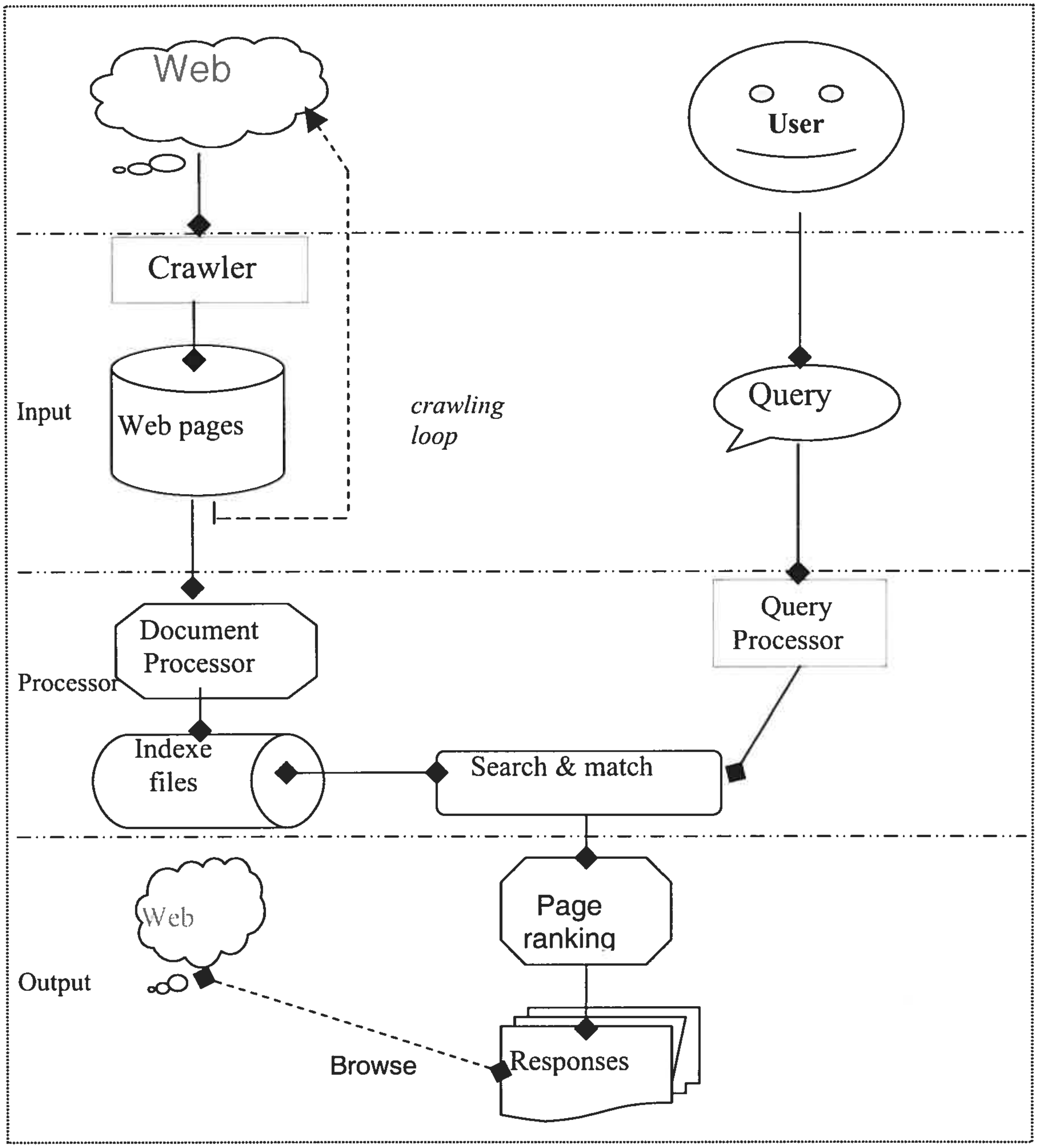


Figure 3.4 *Layers of a web search system*

Chapter 4
Domain specific web search

Web search systems are characterized by their strategies. Our crawler combines several techniques. It starts with a small set of interesting pages obtained by interrogating commercial search engines with some judiciously chosen keywords. The crawler then proceeds to explore the neighbouring web by following links from the initial page set. However, the real problem a targeted crawler must address is the analysis of page content to determine the pertinence of each page and its links. Pages on commercial sites generally follow a standard layout with the interesting data in the center surrounded by generic header and footer information combined with navigation sidebars. In order to intelligently classify pages, the crawler must first do a *structural analysis* to identify the various functional blocks on a page (header, footer, etc…) and extract its information core; then it can go on to a *lexical analysis* of this core to determine whether the page is relevant to the area of interest.

4.1 Proposed architecture

Figure 4.1 shows the general architecture of the system. The initial query will include a city name where from a menu of city plus user selected terms and the results will be used as seeds (initial URLs), the number of seeds is set as a parameter. We start with a small set of web page URLs obtained from commercial search engines such as Google or Yahoo. Then we proceed to explore links from those pages. In order to explore efficiently, we must evaluate the pertinence both of pages and links, and discard irrelevant data. This is underlined by the several places where data is discarded.

The diagram shows that various techniques come into play. After a quick initial evaluation, such as file type etc., the crawler begins to do a real job: the classification of pages. In what follows we consider the classification of pages, tunneling, structure analysis and content analysis.

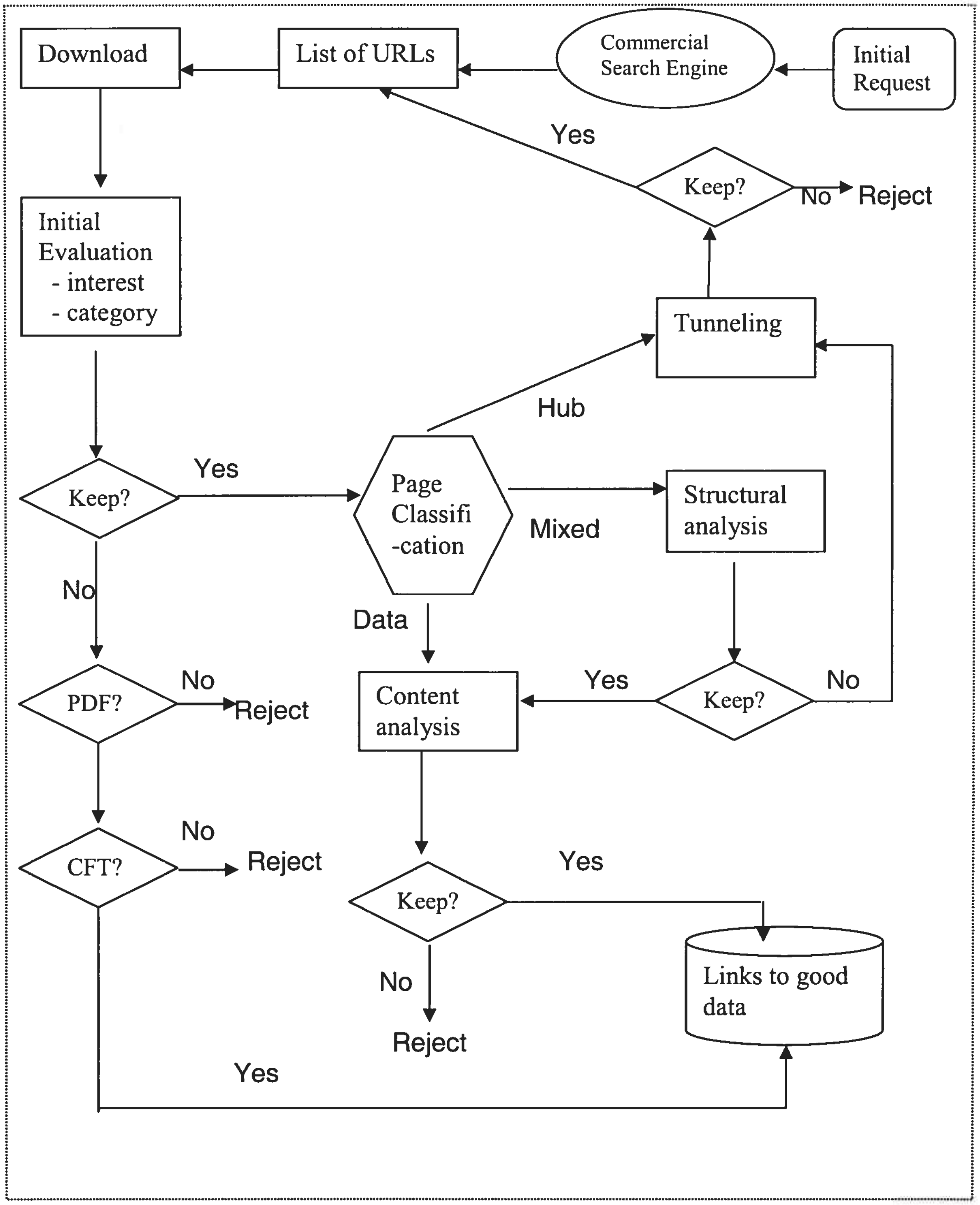
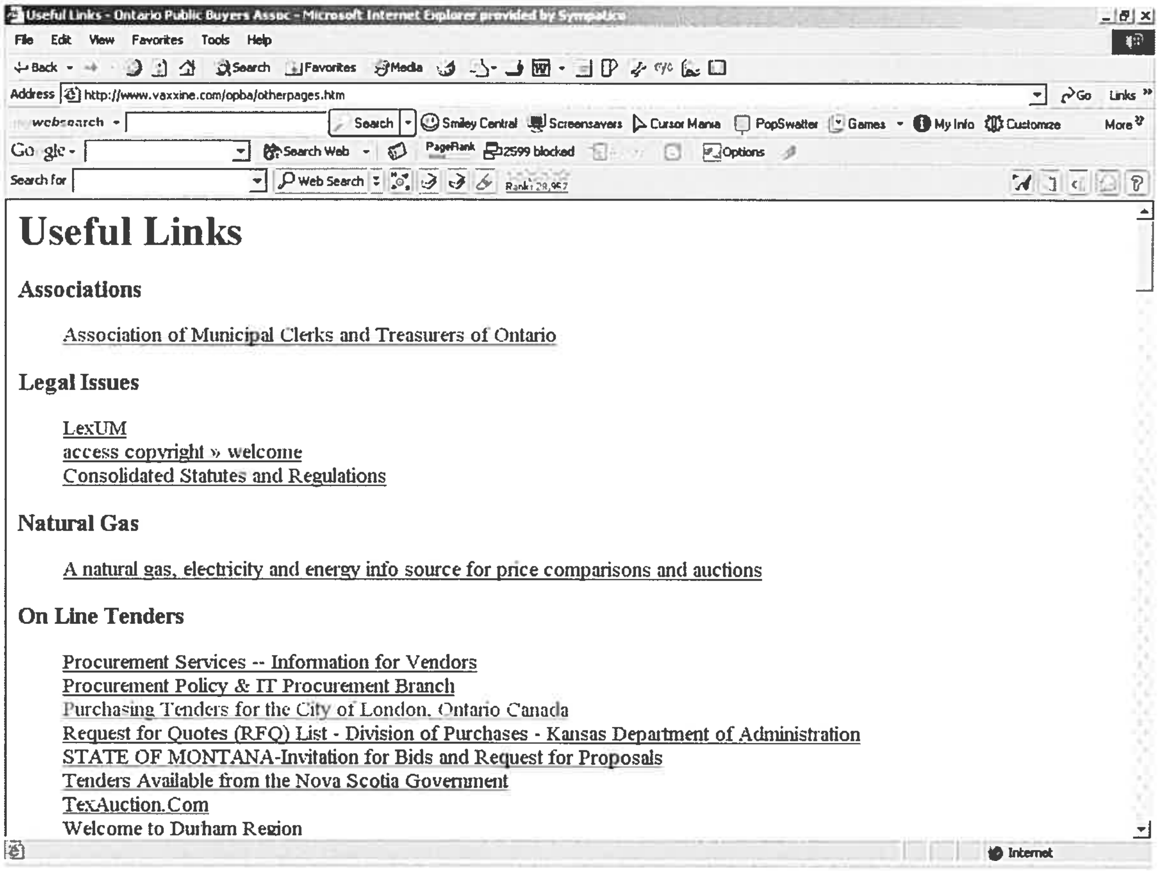


Figure 4.1 Domain specific searching system

4.2 Classification of web pages

To focus on a specific domain, the crawler should pick out links from downloaded pages judiciously. Besides, the particular treatment varies with the web page being crawled. For the different type of web pages, the selection process shouldn't be the same. For some web pages, the crawler may pick out links and crawls further; for some others, it may store or discard them so as to terminate the current crawling process. Normally speaking there are three types of web pages: hub, data and mixed pages. A hub page is a web page on which most of its content are links (see Figure 4.2). A data page is a page on which most of the content is pure text (see Figure 4.3). A mixed page is a web page with a mixture of links and text (see Figure 4.4 type ❶ & Figure 4.5 type❷).


Useful Links - Ontario Public Buyers Assoc - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.vaxxine.com/opba/otherpages.htm
Go Links
websearch Search Smiley Central Screensavers Cursor Mania PopSwatter Games My Info Customize More
Google Search Web PageRank 2599 blocked Options
Search for Web Search
Internet

Useful Links

Associations

Association of Municipal Clerks and Treasurers of Ontario

Legal Issues

LexUM
access copyright » welcome
Consolidated Statutes and Regulations

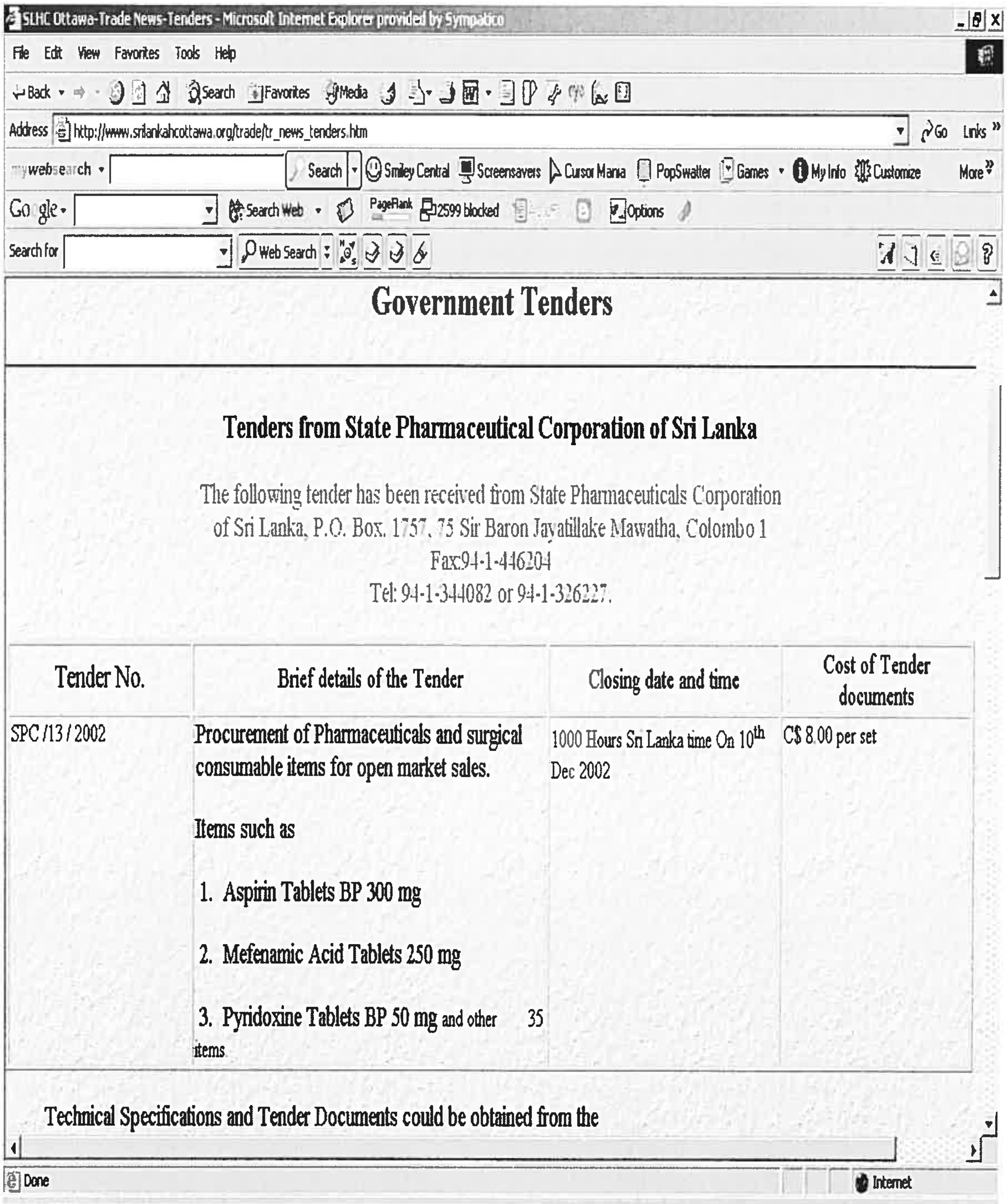
Natural Gas

A natural gas, electricity and energy info source for price comparisons and auctions

On Line Tenders

Procurement Services -- Information for Vendors
Procurement Policy & IT Procurement Branch
Purchasing Tenders for the City of London, Ontario Canada
Request for Quotes (RFQ) List - Division of Purchases - Kansas Department of Administration
STATE OF MONTANA-Invitation for Bids and Request for Proposals
Tenders Available from the Nova Scotia Government
TexAuction.Com
Welcome to Durham Region

Figure 4.2 *Typical "Hub" page*


SLHC Ottawa-Trade News-Tenders - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.srilankahcottawa.org/trade/tr_news_tenders.htm
Go Links
mywebsearch Search Smiley Central Screensavers Cursor Mania PopSwatter Games My Info Customize More
Google Search Web PageRank 2599 blocked Options
Search for Web Search
Done
Internet

Government Tenders

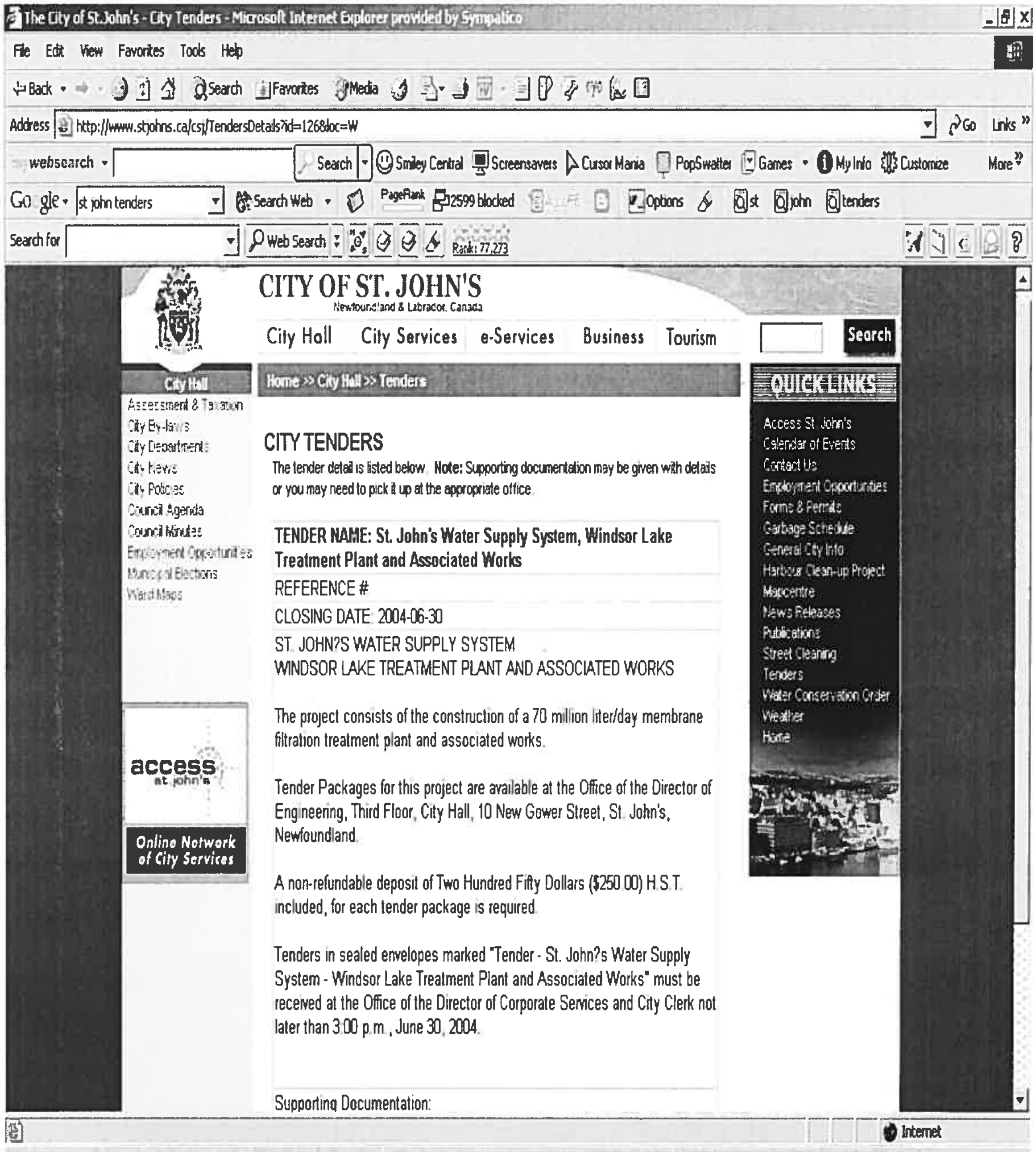
Tenders from State Pharmaceutical Corporation of Sri Lanka

The following tender has been received from State Pharmaceuticals Corporation
of Sri Lanka, P.O. Box. 1757, 75 Sir Baron Jayatillake Mawatha, Colombo 1
Fax:94-1-446204
Tel: 94-1-344082 or 94-1-326227.

Tender No.	Brief details of the Tender	Closing date and time	Cost of Tender documents
SPC /13 / 2002	Procurement of Pharmaceuticals and surgical consumable items for open market sales. Items such as 1. Aspirin Tablets BP 300 mg 2. Mefenamic Acid Tablets 250 mg 3. Pyridoxine Tablets BP 50 mg and other 35 items	1000 Hours Sri Lanka time On 10th Dec 2002	C$ 8.00 per set

Technical Specifications and Tender Documents could be obtained from the

Figure 4.3 *Example of a data page*


The City of St.John's - City Tenders - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.stjohns.ca/csj/TendersDetails?id=126&loc=W
Go Links
websearch Search Smiley Central Screensavers Cursor Mania PopSwatter Games My Info Customize More
Google st john tenders Search Web PageRank 2599 blocked Options st john tenders
Search for Web Search Rank: 77,273
CITY OF ST. JOHN'S
Newfoundland & Labrador, Canada
City Hall City Services e-Services Business Tourism Search
City Hall
Assessment & Taxation
City By-laws
City Departments
City News
City Policies
Council Agenda
Council Minutes
Employment Opportunities
Municipal Elections
Ward Maps
access st. john's
Online Network of City Services
Home >> City Hall >> Tenders
QUICK LINKS
Access St. John's
Calendar of Events
Contact Us
Employment Opportunities
Forms & Permits
Garbage Schedule
General City Info
Harbour Clean-up Project
Mapcentre
News Releases
Publications
Street Cleaning
Tenders
Water Conservation Order
Weather
Home
Internet

CITY TENDERS

The tender detail is listed below. **Note:** Supporting documentation may be given with details or you may need to pick it up at the appropriate office.

TENDER NAME: St. John's Water Supply System, Windsor Lake Treatment Plant and Associated Works

REFERENCE #

CLOSING DATE: 2004-06-30

ST. JOHN?S WATER SUPPLY SYSTEM
WINDSOR LAKE TREATMENT PLANT AND ASSOCIATED WORKS

The project consists of the construction of a 70 million liter/day membrane filtration treatment plant and associated works.

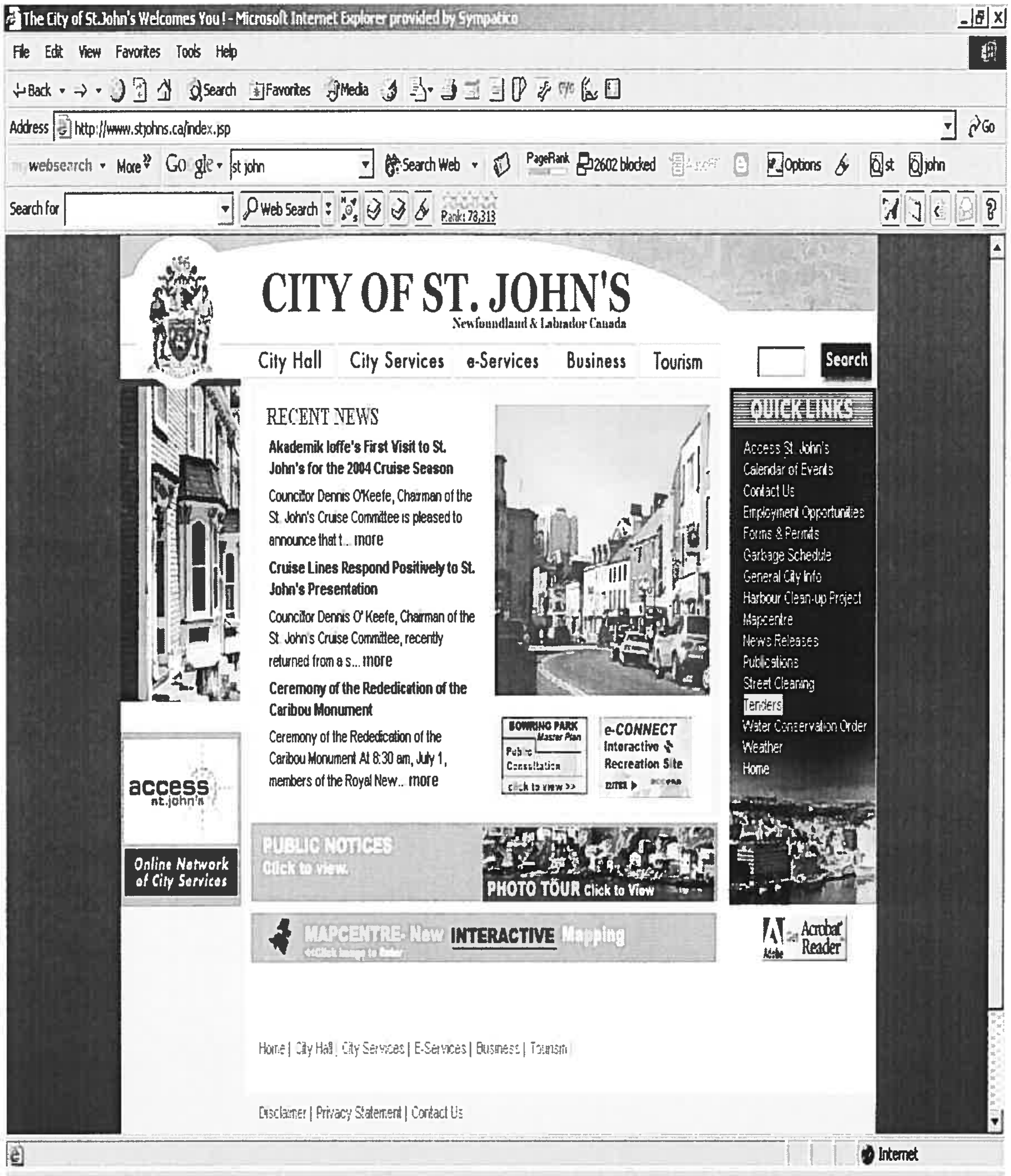
Tender Packages for this project are available at the Office of the Director of Engineering, Third Floor, City Hall, 10 New Gower Street, St. John's, Newfoundland.

A non-refundable deposit of Two Hundred Fifty Dollars ($250.00) H.S.T. included, for each tender package is required.

Tenders in sealed envelopes marked "Tender - St. John?s Water Supply System - Windsor Lake Treatment Plant and Associated Works" must be received at the Office of the Director of Corporate Services and City Clerk not later than 3:00 p.m., June 30, 2004.

Supporting Documentation:

Figure 4.4 *A "Mixed" web page (type ❶)*



The City of St.John's Welcomes You! - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.stjohns.ca/index.jsp
Go
websearch More Google st john Search Web PageRank 2602 blocked Options st john
Search for Web Search Rank: 78,313
CITY OF ST. JOHN'S
Newfoundland & Labrador Canada
City Hall
City Services
e-Services
Business
Tourism
Search
QUICK LINKS
Access St. John's
Calendar of Events
Contact Us
Employment Opportunities
Forms & Permits
Garbage Schedule
General City Info
Harbour Clean-up Project
Mapcentre
News Releases
Publications
Street Cleaning
Tenders
Water Conservation Order
Weather
Home
access st.john's
Online Network of City Services
BOWRING PARK Master Plan
Public Consultation
click to view >>
e-CONNECT
Interactive Recreation Site
PUBLIC NOTICES
Click to view.
PHOTO TOUR Click to View
MAPCENTRE- New INTERACTIVE Mapping
<<Click image to Enter
Acrobat Reader
Home | City Hall | City Services | E-Services | Business | Tourism
Disclaimer | Privacy Statement | Contact Us
Internet

RECENT NEWS

Akademik Ioffe's First Visit to St. John's for the 2004 Cruise Season

Councillor Dennis O'Keefe, Chairman of the St. John's Cruise Committee is pleased to announce that t... more

Cruise Lines Respond Positively to St. John's Presentation

Councillor Dennis O' Keefe, Chairman of the St. John's Cruise Committee, recently returned from a s... more

Ceremony of the Rededication of the Caribou Monument

Ceremony of the Rededication of the Caribou Monument At 8:30 am, July 1, members of the Royal New... more

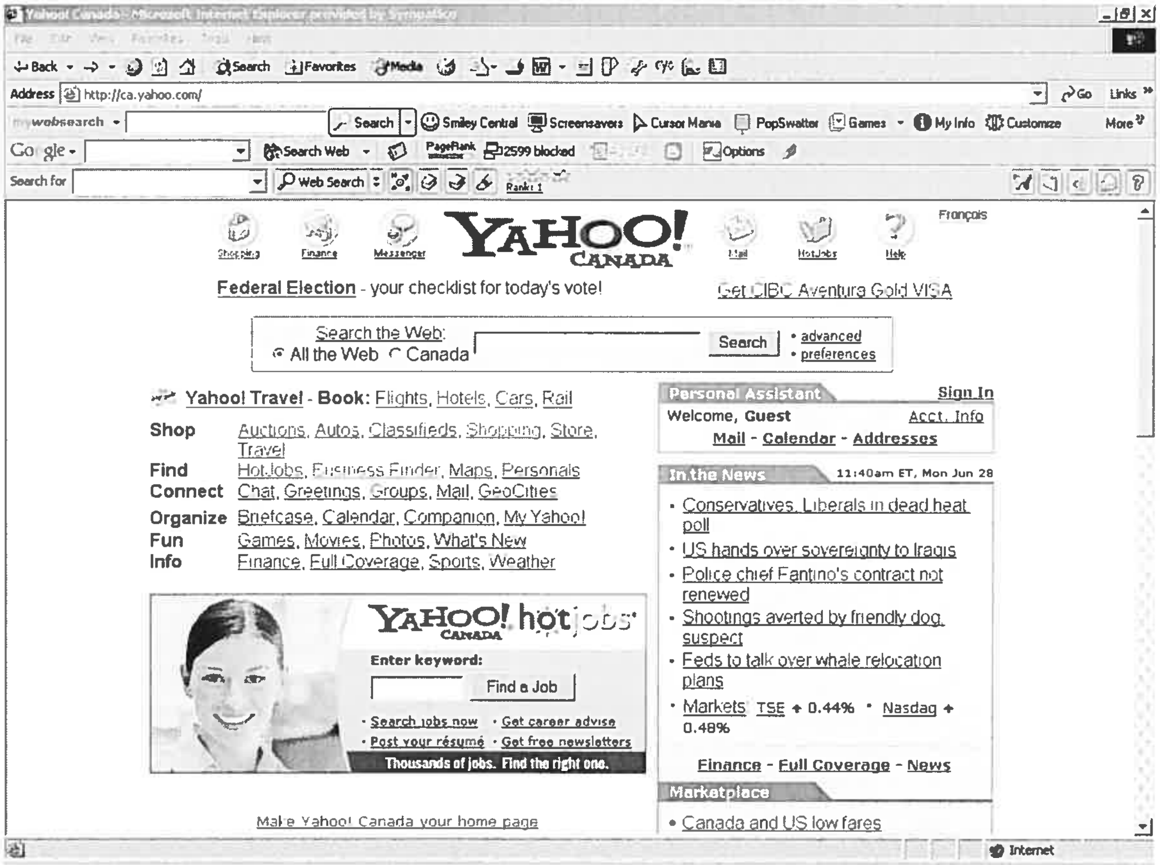
Figure 4.5 *A "Mixed" web page (type ❷)*

Type ❶ pages can be treated as a data page, whereas type ❷ pages are rather like hub pages. This is because each piece of text in type ❷ page is a brief introduction for the nearest link. Although it is not easily detected by a crawler, the ratio of links/strings is useful in differentiating between type ❶ and type ❷ pages.

4.3 Tunneling

When a crawler explores a hub page, it will analyze anchor words of all the links within the page so as to explore some links further with close meaning of the query term. Sometimes, however, the anchor words in relevant links do not match any in the query.

For example, if we search "tender" on the following hub page (see the figure 4.6),



Federal Election - your checklist for today's vote!

Get CIBC Aventura Gold VISA

Search the Web:
● All the Web ○ Canada [Search]
• advanced
• preferences

Yahoo! Travel - **Book:** Flights, Hotels, Cars, Rail

Shop Auctions, Autos, Classifieds, Shopping, Store, Travel
Find HotJobs, Business Finder, Maps, Personals
Connect Chat, Greetings, Groups, Mail, GeoCities
Organize Briefcase, Calendar, Companion, My Yahoo!
Fun Games, Movies, Photos, What's New
Info Finance, Full Coverage, Sports, Weather

YAHOO! CANADA hotjobs
Enter keyword:
[Find a Job]
• Search jobs now • Get career advise
• Post your résumé • Get free newsletters
Thousands of jobs. Find the right one.

Make Yahoo! Canada your home page

Personal Assistant **Sign In**
Welcome, Guest Acct. Info
Mail - **Calendar** - **Addresses**

In the News 11:40am ET, Mon Jun 28

- Conservatives, Liberals in dead heat poll
- US hands over sovereignty to Iraqis
- Police chief Fantino's contract not renewed
- Shootings averted by friendly dog, suspect
- Feds to talk over whale relocation plans
- Markets TSE + 0.44% • Nasdaq + 0.48%

Finance - **Full Coverage** - **News**

Marketplace

- Canada and US low fares

Figure 4.6 *Tunneling*

we can't directly find any anchor word matching the term "tender", but we should explore some pertinent links like "Shopping" or "Business Finder" first rather than exploring some other irrelevant links or even discard this hub page.

The phenomenon that a crawler has to pass through irrelevant pages whose content does not match exactly on query in order to reach useful pages is called tunneling (Figure 4.6). Tunneling often occurs in hub pages. In this case, the crawler needs to adjust whatever widening the matching criteria or extending the query term to tackle the tunneling [Bergmark et al. 2002].

Another example of tunneling is when the hub pages provide "search" function. The crawler can auto-fill "search" form with the query term, send a requirement, and get a response page to avoid tunneling.

4.4 Web page structural analysis

Text data, hyperlinks, images are organically placed inside web pages. An explicit visual structure of a web page can help us to understand it semantically. Such structure can be partly used in web pages' classification since an analysis of the structure may reveal logical relevancy between sections in the web page.

Web page structural analysis combines both structural and semantics information of a page in order to get to a logical page schema.

4.4.1 Web page structure

Semantic information of a web page often needs to be organized with various sections so as to represent author's view or linkage relation. This structural requirement reflects in both coding html tags and arranging logical content [Carchiolo et al. 2003].

(1) HTML structure

Tags are used to organize blocks of text representing the same information. They also control graphical appearance of a web page. Tags like <table>, <p>, <hr>, <map>, <a>, <frame>, <td>, <tr>, <option>, <area>, <li> often present a logical section or a subsection in a web page. Besides some properties play an important role in a semantic partition of a web page, such as <td bgcolor =>, <front size=>, <td width=>.

As mentioned above, although a lot of html tags can represent relative independent logical meaning in a web page, the <table> tag is better for analyzing. In a correctly formatted web page, various logical sections are located between opening tags (i.e. <table>) and closing tags (i.e. </table>). All the <table> tags of a web page can be constructed in a tree structure. Analyzing this tree can helps in recognizing logical structure of a web page.

(2) Logical structure

Often the HTML structure reflects the logical organization of a web page. Just as written text follows certain conventions for example starting with a title page, then an introduction, a body and a conclusion followed by references; the constituent parts of a web page often fall in certain typical categories.

Generally a web page can be partitioned as several logical sections:

1) a *document information section:* metadata concerning the document, as document type specified in <!doctype>, referring DTD, author etc.

2) a *logical heading:* the name, title or the logo of the site, usually placed at the beginning of the document. Denoted by the tag <head>.

3) a *logical footer:* generally placed at the bottom of page and whose content is often a mail link to webmasters or text with copyright, privacy etc.

4) *logical links:* a set of HTML tags representing a link towards another page. Denoted by the tag <a href>.

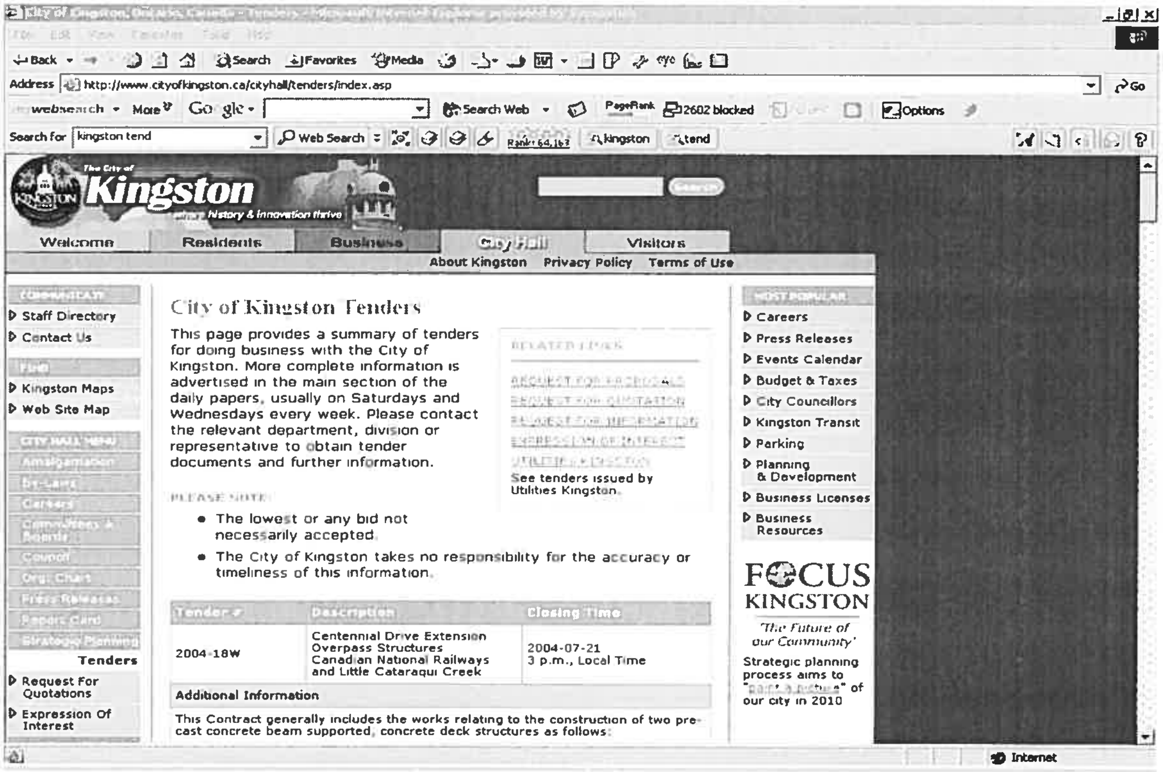
5) an *index:* a group of logical links having similar characteristics and graphical appearance. Denoted by the tag <a href>.

6) *logical data:* containing raw data --- information core of the page

7) *interactive sections:* forms where users interact with the page; this section generally includes codes such as cgi, php, applets or script, whose meaning is hard to understand. This section often contains <form>, <input> or <select>. It is easily to be identified or automatic filled by a piece of program.

4.4.2 Analyzing structural information

Not all the web pages require structural analysis; only those pages with a relative complex structure, similar to Figure 4.4 or Figure 4.7. These web pages normally have the logical section 1) to 6) or 1) to 7) mentioned above. Since most logical data sections mainly contain raw data, as a simplified example of using structural analysis, we extract the logical data section from these web pages for lexical analysis. To do the structural analysis, we decompose a web page owning a complex structure and being composed with a lot of tables to a set of sub-tables, analyze each of them, and then pick out the proper one --- the logical data section of the web page. When a crawler explores a mixed page, it will analyze the structure and content of that page. The crawler will extract the main content of that page and then do content analysis. For example, see Figure 4.7, Figure 4.8.



City of Kingston Tenders

This page provides a summary of tenders for doing business with the City of Kingston. More complete information is advertised in the main section of the daily papers, usually on Saturdays and Wednesdays every week. Please contact the relevant department, division or representative to obtain tender documents and further information.

RELATED LINKS

REQUEST FOR PROPOSALS
REQUEST FOR QUOTATION
REQUEST FOR INFORMATION
EXPRESSION OF INTEREST
UTILITIES KINGSTON
See tenders issued by Utilities Kingston.

PLEASE NOTE:

- The lowest or any bid not necessarily accepted.
- The City of Kingston takes no responsibility for the accuracy or timeliness of this information.

Tender #	Description	Closing Time
2004-18W	Centennial Drive Extension Overpass Structures Canadian National Railways and Little Cataraqui Creek	2004-07-21 3 p.m., Local Time

Additional Information

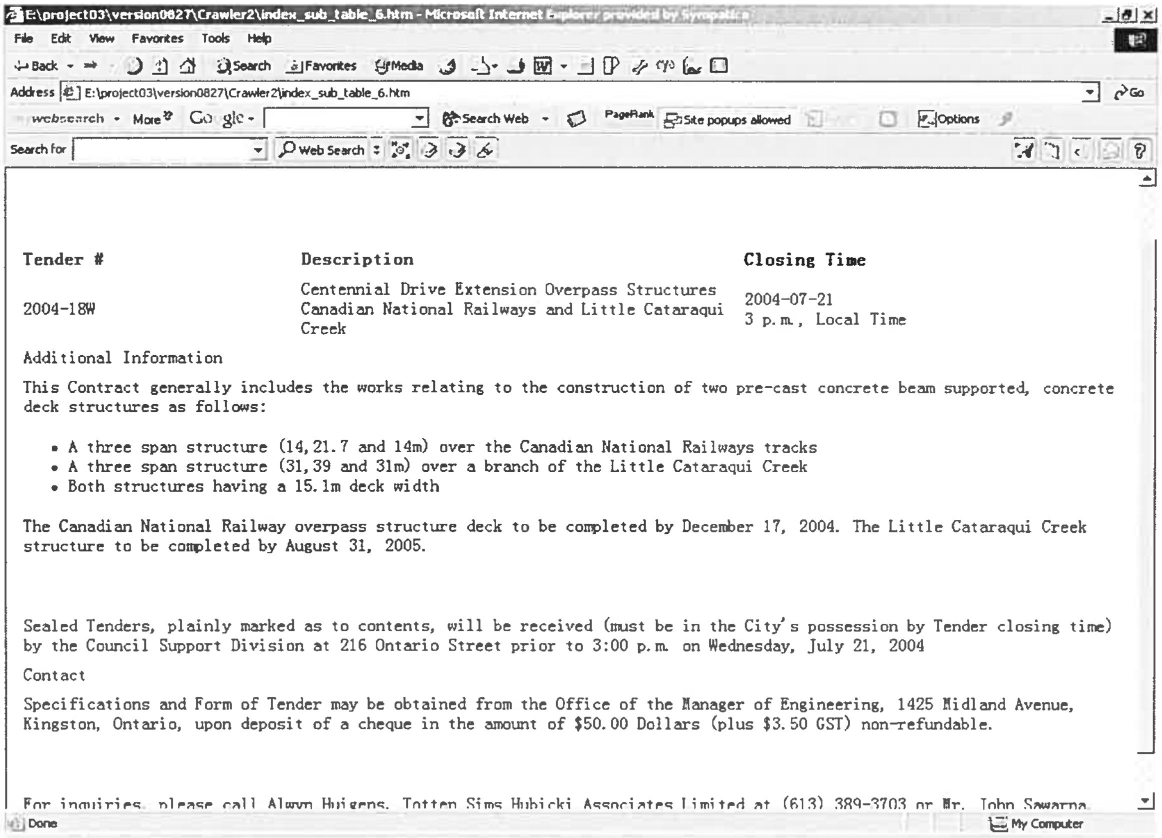
This Contract generally includes the works relating to the construction of two pre-cast concrete beam supported, concrete deck structures as follows:

Figure 4.7 *"Mixed" page before data extraction*

The analyzing process decomposes a mixed page with a lot of imbricated tables, then counts links and strings of each table, calculates a ratio (number of links/ number of strings) and a value for each table.

The value degree = "tender" occurring frequency / ratio, if ratio not equal zero; otherwise indicate an experimental value.

A table with the biggest value degree is extracted out.


E:\project03\version0827\Crawler2\index_sub_table_6.htm - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address E:\project03\version0827\Crawler2\index_sub_table_6.htm
Go
websearch More Google Search Web PageRank Site popups allowed Options
Search for Web Search
Done
My Computer

Tender #	Description	Closing Time
2004-18W	Centennial Drive Extension Overpass Structures Canadian National Railways and Little Cataraqui Creek	2004-07-21 3 p.m., Local Time

Additional Information

This Contract generally includes the works relating to the construction of two pre-cast concrete beam supported, concrete deck structures as follows:

- A three span structure (14,21.7 and 14m) over the Canadian National Railways tracks
- A three span structure (31,39 and 31m) over a branch of the Little Cataraqui Creek
- Both structures having a 15.1m deck width

The Canadian National Railway overpass structure deck to be completed by December 17, 2004. The Little Cataraqui Creek structure to be completed by August 31, 2005.

Sealed Tenders, plainly marked as to contents, will be received (must be in the City's possession by Tender closing time) by the Council Support Division at 216 Ontario Street prior to 3:00 p.m. on Wednesday, July 21, 2004

Contact

Specifications and Form of Tender may be obtained from the Office of the Manager of Engineering, 1425 Midland Avenue, Kingston, Ontario, upon deposit of a cheque in the amount of $50.00 Dollars (plus $3.50 GST) non-refundable.

For inquiries, please call Alwyn Huigens, Totten Sims Hubicki Associates Limited at (613) 389-3703 or Mr. John Sawarna

Figure 4.8 *"Mixed" page after extraction*

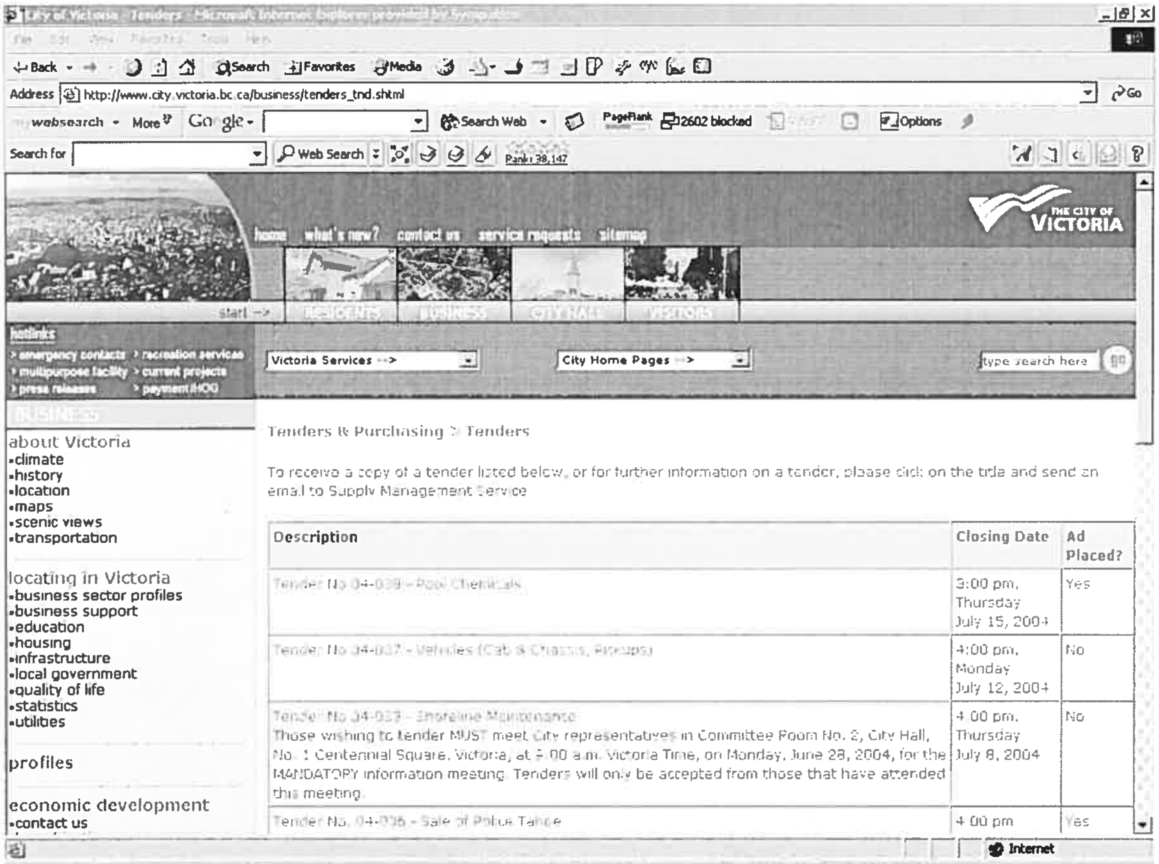
We call the main content part extract from the "mixed" web page an information core. If the information core matching a primary requirement of querying, for example, it contains the term "tender", then the page where the information core locates will be sent to do content analysis for further evaluation. Otherwise the page will be classified as a hub page (see figure 4.5).

4.5 Web page content lexical analysis

Lexical analysis is to classify web pages according to their content. Lexical analysis will transform the web pages into a representation suitable for classification task. It includes the following steps: collecting a training set, extracting features of it, calculating similarity, classifying the web page automatically.

4.5.1 Collecting a training set

For the domain specific search, a training set relate to a specific domain. To collect the training set, a set of related web pages (HTML files) should be downloaded firstly. After removing HTML tags and stop-words from those HTML files, (if necessary, perform word stemming) the set of web pages is transformed to a collection of documents. In the "call for tender" domain, different sites have various written formats, hence collecting a set of tenders from several cities is necessary to extract representation of tenders. The examples of various tenders are shown as following (Figure 4.9, 4.10, 4.11 & 4.12).



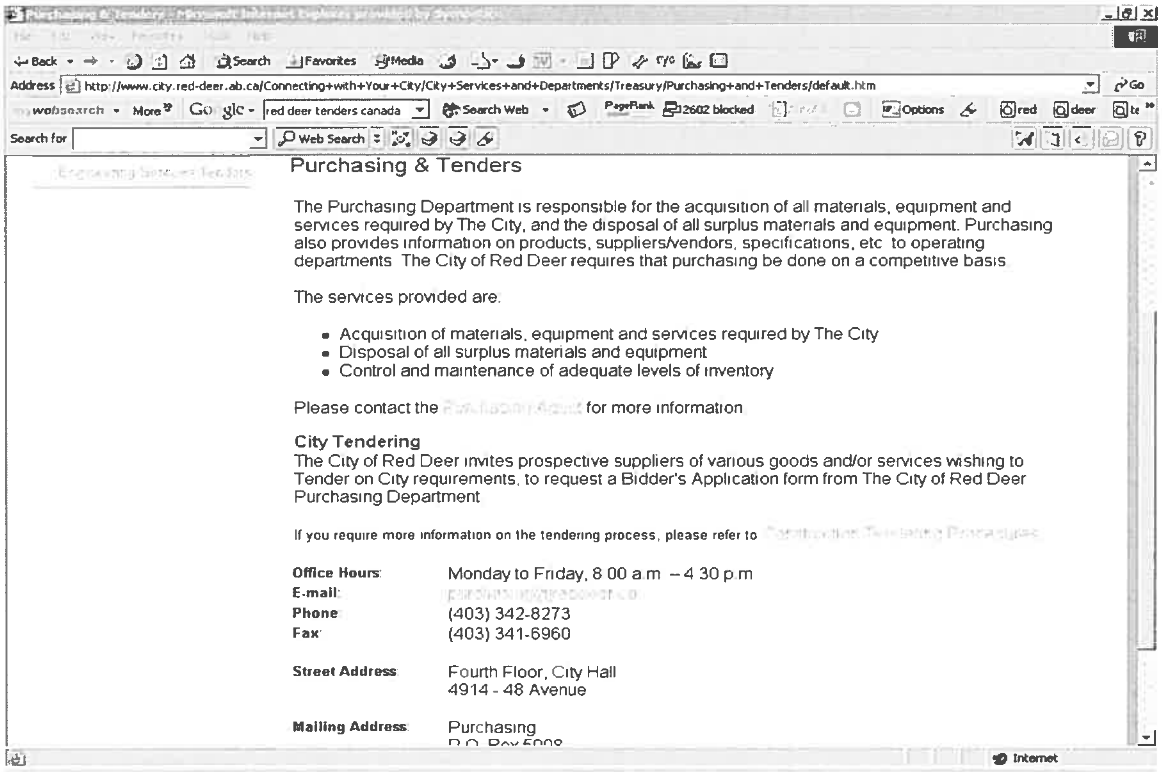
Address http://www.city.victoria.bc.ca/business/tenders_tnd.shtml

Tenders & Purchasing > Tenders

To receive a copy of a tender listed below, or for further information on a tender, please click on the title and send an email to Supply Management Service

Description	Closing Date	Ad Placed?
Tender No 04-039 - Pool Chemicals	3:00 pm, Thursday July 15, 2004	Yes
Tender No 04-037 - Vehicles (Cab & Chassis, Pickups)	4:00 pm, Monday July 12, 2004	No
Tender No 04-033 - Shoreline Maintenance Those wishing to tender MUST meet City representatives in Committee Room No. 2, City Hall, No. 1 Centennial Square, Victoria, at 9:00 a.m. Victoria Time, on Monday, June 28, 2004, for the MANDATORY information meeting. Tenders will only be accepted from those that have attended this meeting.	4:00 pm, Thursday July 8, 2004	No
Tender No. 04-036 - Sale of Police Tahoe	4:00 pm	Yes

Figure 4.9 *City of Victoria, BC*


Back
Search
Favorites
Media
Address
http://www.city.red-deer.ab.ca/Connecting+with+Your+City/City+Services+and+Departments/Treasury/Purchasing+and+Tenders/default.htm
Go
websearch
More
Google
red deer tenders canada
Search Web
PageRank
2602 blocked
Options
red
deer
te
Search for
Web Search
Engineering Services Tenders
Internet

Purchasing & Tenders

The Purchasing Department is responsible for the acquisition of all materials, equipment and services required by The City, and the disposal of all surplus materials and equipment. Purchasing also provides information on products, suppliers/vendors, specifications, etc to operating departments The City of Red Deer requires that purchasing be done on a competitive basis

The services provided are:

- Acquisition of materials, equipment and services required by The City
- Disposal of all surplus materials and equipment
- Control and maintenance of adequate levels of inventory

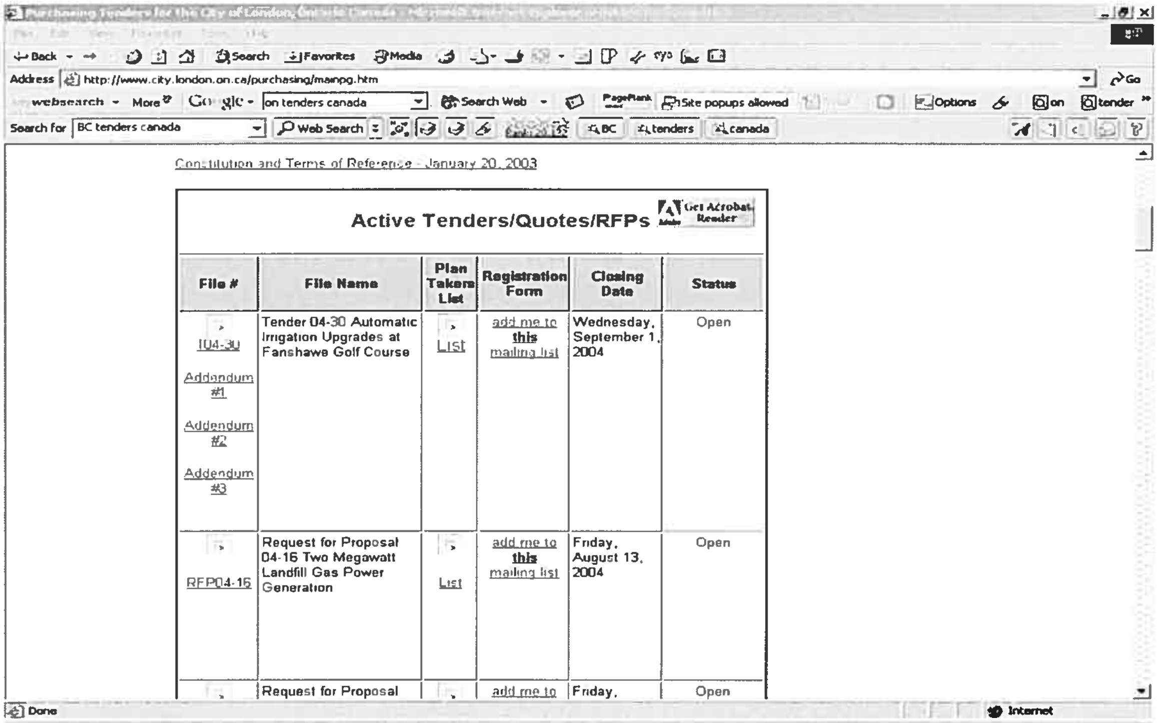
Please contact the Purchasing Agent for more information

City Tendering
The City of Red Deer invites prospective suppliers of various goods and/or services wishing to Tender on City requirements, to request a Bidder's Application form from The City of Red Deer Purchasing Department

If you require more information on the tendering process, please refer to Construction Tendering Procedures

Office Hours:	Monday to Friday, 8 00 a.m – 4 30 p.m
E-mail:	[illegible]
Phone:	(403) 342-8273
Fax:	(403) 341-6960
Street Address:	Fourth Floor, City Hall 4914 - 48 Avenue
Mailing Address:	Purchasing P.O. Box 5008

Figure 4.10 *City of Red deer, Alberta*

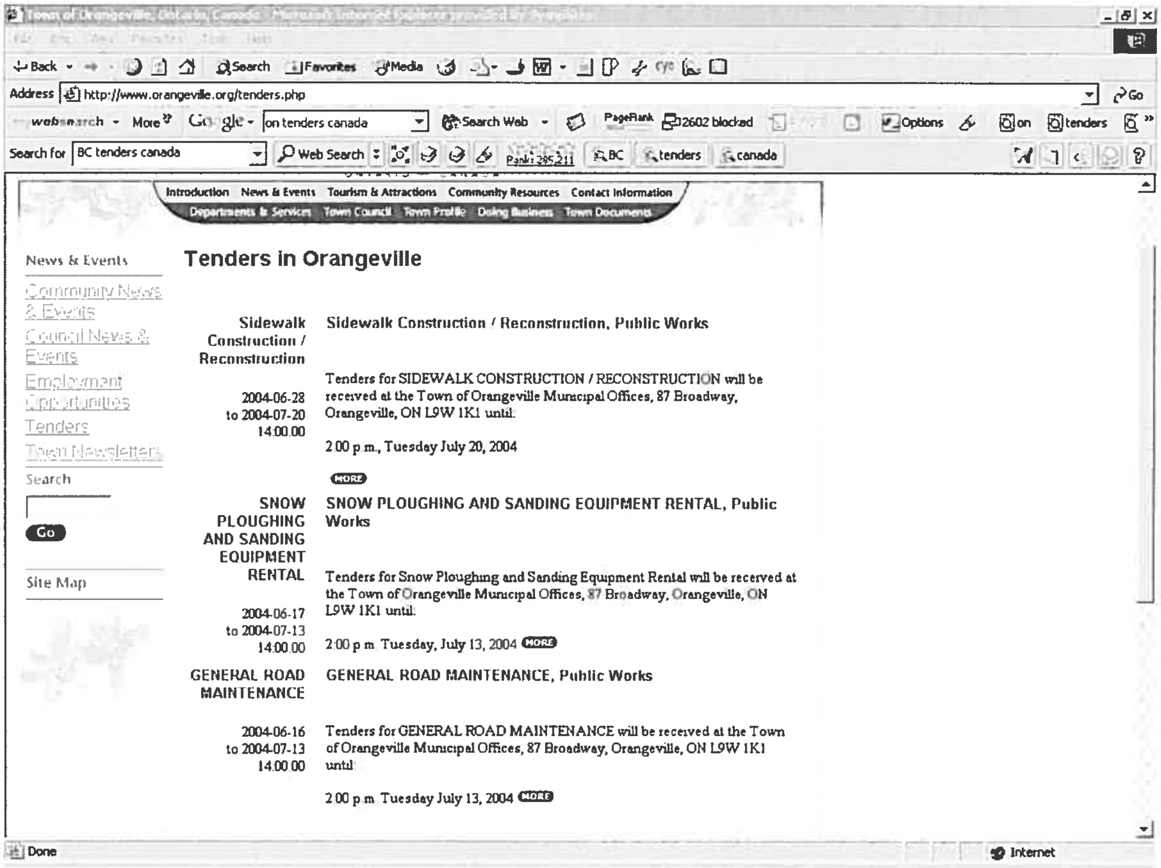


Constitution and Terms of Reference - January 20, 2003

Active Tenders/Quotes/RFPs

File #	File Name	Plan Takers List	Registration Form	Closing Date	Status
T04-30 Addendum #1 Addendum #2 Addendum #3	Tender 04-30 Automatic Irrigation Upgrades at Fanshawe Golf Course	List	add me to **this** mailing list	Wednesday, September 1, 2004	Open
RFP04-16	Request for Proposal 04-16 Two Megawatt Landfill Gas Power Generation	List	add me to **this** mailing list	Friday, August 13, 2004	Open
	Request for Proposal		add me to	Friday,	Open

Figure 4.11 *City of London, Ontario*



Tenders in Orangeville

Sidewalk Construction / Reconstruction	Sidewalk Construction / Reconstruction, Public Works
2004-06-28 to 2004-07-20 14:00:00	Tenders for SIDEWALK CONSTRUCTION / RECONSTRUCTION will be received at the Town of Orangeville Municipal Offices, 87 Broadway, Orangeville, ON L9W 1K1 until: 2:00 p.m., Tuesday July 20, 2004 MORE
SNOW PLOUGHING AND SANDING EQUIPMENT RENTAL	SNOW PLOUGHING AND SANDING EQUIPMENT RENTAL, Public Works
2004-06-17 to 2004-07-13 14:00:00	Tenders for Snow Ploughing and Sanding Equipment Rental will be received at the Town of Orangeville Municipal Offices, 87 Broadway, Orangeville, ON L9W 1K1 until: 2:00 p.m. Tuesday, July 13, 2004 MORE
GENERAL ROAD MAINTENANCE	GENERAL ROAD MAINTENANCE, Public Works
2004-06-16 to 2004-07-13 14:00:00	Tenders for GENERAL ROAD MAINTENANCE will be received at the Town of Orangeville Municipal Offices, 87 Broadway, Orangeville, ON L9W 1K1 until: 2:00 p.m. Tuesday July 13, 2004 MORE

Figure 4.12 *Town of Orangeville, Ontario*

4.5.2 Feature extraction

From the Figure 4.9 to 4.12, we can see various formats of tenders. After collecting a set of cities' tenders, we need to extract their common features. The most commonly used document representation is the so-called vector space model. In this model, documents are represented by vectors of words. A collection of documents which is represented by a word-by-document matrix **A**, where each entry represents the occurrences of a word in a document, i.e., $\mathbf{A} = (\mathbf{a}_{ik})$, where $\mathbf{a}_{ik}$ is the weight of word i in document k. These techniques are well known in the IR fields, for a further overview, see [Aas et al. 1999].

There are several ways of determining the weight $\mathbf{a}_{ik}$, most of the approaches are based on two empirical observations regarding text:

1) the more times a word occurs in a document, the more relevant it is to the topic of the document.
2) the more times the word occurs throughout all documents in the collection, the more poorly it discriminates between documents.

Suppose we have the statistic matrix A0 derive from the collection:

$$A0 = \begin{array}{l} \text{word}_1 \\ \text{word}_2 \\ \ldots \\ \text{word}_M \end{array} \left(\begin{array}{c} f_{11}, f_{12}, \ldots f_{1N} \\ f_{21}, f_{22}, \ldots f_{2N} \\ \ldots \\ f_{M1}, f_{M2}, \ldots f_{MN} \end{array}\right)$$

let f_{ik} be the frequency of word i in document k, M number of words in collection N number of documents in collection, n_i is the total number of times word i occurs in the whole collection.

If adopting the tf/idf weighting, $\mathbf{a}_{ik} = f_{ik} * log(N/ n_i)$, thus we got the matrix **A:**

$$A = (\mathbf{a}_{ik}) = \begin{array}{l} \text{word}_1 \\ \text{word}_2 \\ \ldots \\ \text{word}_M \end{array} \left(\begin{array}{c} \mathbf{a}_{11}, \mathbf{a}_{12}, \ldots \mathbf{a}_{1N} \\ \mathbf{a}_{21}, \mathbf{a}_{22}, \ldots \mathbf{a}_{2N} \\ \ldots \\ \mathbf{a}_{M1}, \mathbf{a}_{M2}, \ldots \mathbf{a}_{MN} \end{array}\right)$$

There may be a lot of words in the collection. Hence, the dimensionality of the feature space (matrix A) is very high. We can remove non-informative words from documents through the feature selection so as to improve effectiveness and reduce computational complexity.

The document frequency for a word is the number of documents in which the word occurs. The document frequency threshold can be chosen as the feature selection method to reduce the original feature set.

Applying the main idea above mentioned, we collect a set of tenders in 10 cities and do lexical analysis as following:

1) http://www.oshawa.ca/cit_hall/tenders.asp
2) http://www.city.brampton.on.ca/purchasing/rfp_2004-057.tml
3) http://www.city.brampton.on.ca/purchasing/rfp_2004-040.tml
4) http://www.city.cambridge.on.ca/cs_corporate/purchasing_tenders_list.php
5) http://www.city.kitchener.on.ca/tenders/tender.asp
6) http://www.city.london.on.ca/Purchasing/mainpg.htm
7) http://www.city.niagarafalls.on.ca/cityhall/qtenders.html
8) http://www.region.peel.on.ca/finance/purchasing/biddocs/index.htm
9) http://www.city.greatersudbury.on.ca/pubapps/tenders/
10) http://www.city.toronto.on.ca/tenders/tenders_to.htm
11)http://www.region.waterloo.on.ca/web/region.nsf/97dfc347666efede85256e590071a3d4/4
a3823e3515da81e85256f11004b50c4!OpenDocument

Words whose occurring frequency larger than 10 are:

Frequency	Word	Frequency	Word
133	click	35	mailing
83	tender	34	date
85	list	34	proposal
76	acrobat	33	opening
75	format	31	document
74	pending	30	purchasing
63	ad	26	city
62	services	24	supply
40	contract	24	tenders
40	addendum	23	bid
40	request	23	issued
38	september	22	sep
35	contact	22	peel
35	add	20	information
34	august	20	closed

Frequency	Word	Frequency	Word
20	works	12	proposals
19	wednesday	12	division
19	aug	12	road
19	extended	12	bids
17	rfp	12	site
16	quotation	12	Brampton
16	plan	12	hall
16	council	12	perform
16	emergency	12	takers
16	construction	12	links
15	street	12	view
14	deposit	11	public
14	form	11	pdf
14	friday	11	opportunities
14	replacement	11	regional
14	awarded	11	removal
14	search	11	community
13	download	11	register
13	open	10	west
13	transportation	10	centre
13	closing	10	delivery
13	buyer	10	full
13	home	10	floor
12	sept	10	fill

Table 4.1 Word Frequency

From above results we can see that CFT content lexical analysis is important but importance of a word doesn't match exactly with its occurring frequency in a CFT document. For example, the first 8 high-frequency words are: click (133), list (85), tender (83), acrobat (76), format (75), pending (74), ad (63), services (62). If we evaluate a document with these nine words, we can hardly say the document is a CFT or not. On the contrary, some words play a very important role in a CFT document, though its occurring frequency is not too high. For example, occurring frequency of word "closing" has only thirteen. In fact, the word "closing" appears all in content of URL 1) to URL10) as an important sub-title "closing date". Therefore CFT feature words should consider both CFT lexical analysis and CFT format (structure analysis). Different word has different tender format weight in a CFT document. From format point of view, a CFT normally has following features:

1) *token words*: tender(83), tenders(24), bid(23), bids(12), rfp(17), quotation (16), solicitation(2), proposal(12) etc.

2) *time tag*: closing(13), closed(20), opening(33), open(13), pending(74),date(34), issued(23) etc.

3) *action words*: request(40), contract(40),contact(35), purchasing(30), delivery(10) etc.
4) *requirements*: document(31), addendum(40), information (20)
5) *address words*: add. (35), city (26), street(15), road(12), mailing(35)
6) *value words*: deposit(14), awarded(13), buyer(13)

Although the basic CFT format is the same, the words above may vary with different area or different speciality. The most typical features of CFT are token words and time tag. As long as we combine these two features, the precision of query increase a lot. For example, suppose we want to search CFT in the city of Peel. A quick experiment is to query Google with terms "peel tender", "peel tender closing" and "peel tender request":

Query "peel tender":

1) www.taunton.com/finecooking/pages/c00161.asp talking about cooking eggplant
2) www.baking911.com/howto_blanch.htm talking about cooking
3) home.comcast.net/~iasmin/ mkcc/MKCCfiles/candiedcitrus**peel**.html talking about cooking
4) www.extension.umn.edu/distribution/ nutrition/components/0555%5Bt01%5D.html cooking
5) www.nyapplecountry.com/edtastehowto.htm cooking
6) www.evergreen.edu/biophysics/ technotes/home/dry_food.htm cooking Asparagus
7) www.healthgoods.com/Education/Nutrition_Information/ Drying_Food/pretreating_vegetables.htm cooking mushrooms
8) www.cooks.com/rec/search/0,1-00,baked_ sweet_potatoes_cinnamon,FF.html cooking potatoes
9) www.wholehealthmd.com/refshelf/ foods_view/1,1523,29,00.html cooking Parsnips
10) www.geocities.com/Heartland/Acres/1012/marpre.html cooking

Query “peel tender closing”:

1) www.region.**peel**.on.ca/housing/**peel**_living/ Peel_Living_Minutes/1980s/1980/plmin19801030.htm a CFT in region Peel

2) www.spc.lk/seventhTender2.html a CFT in Sri Lanka

3) purchasing.**peel**schools.org/documents/2601t.pdf a CFT in region Peel

4) purchasing.**peel**schools.org/documents/2469t.pdf a CFT in region Peel

5)www.nationalparks.nsw.gov.au/npws.nsf/Content/ Tender+advert+roads+fire+trial+maintenance+and+fire+suppression a CFT in region Peel

6) www.pakwatan.com/main/businessnew/ **tender**international.php3 a CFT

7) www.pakwatan.com/main/businessnew/**tender**int.php3 a CFT

8)https://www.raqsb.mto.gov.on.ca/raqs_contractor/ raqscont.nsf/0/25EF082A2079207D85256A1B00626DAF?Opendocument a CFT in region Peel

9)www.city.pg.bc.ca/city_services/supply/2003_documents/T03-17_Construction_Services_Roof_Repairs.pdf a CFT in region Peel

10) www.dpcdsb.org/purchasing/program.cfm a CFT

Query “peel tender request”:

1) www.region.**peel**.on.ca/housing/**peel**_living/ Peel_Living_Minutes/1980s/1980/plmin19801030.htm Bussiness meeting in region Peel

2) purchasing.**peel**schools.org/documents/2589t.pdf a CFT in region Peel

3) www.ipc.on.ca/scripts/index_.asp?action=31& P_ID=15033&N_ID=1&PT_ID=759&U_ID=0 An appeal in Peel

4) www.thriftyfun.com/readers**request**/tf876797_rea.html Cooking

5) www.extension.umn.edu/distribution/ nutrition/components/0555%5Bt01%5D.html Vegetables

6) starbulletin.com/97/07/30/features/**request**.html Fruit

7) www.region.**peel**.on.ca/housing/**peel**_living/ Peel_Living_Minutes/1990s/1998/plmin19980813.htm A Meeting in region Peel

8) www.region.**peel**.on.ca/housing/**peel**_living/ Peel_Living_Minutes/1980s/1984/plmin19841204.htm A Meeting in region Peel

9) www.ext.nodak.edu/extpubs/yf/foods/he187w.htm Vegetables

10) 01wholesale.com/ import-export-directory--201-Import-export---wood.html
International trade

From this quick experiment, we can see that there is huge weight difference between CFT feature words. We adopt three CFT feature lists in our domain specific system to evaluate a web page. The first is a CFT token words list, including “tender”, “bid”, “RFP”, “RFQ”, “solicitation”. The second is a CFT time tag list, including “closing”, “open”, “closed”, “issued”, “date”. And the last is a hybrid CFT list, including “document”, “information”, “contract”, “contact”, “specification”. If a web page doesn’t contain one word in the first list, it is not a CFT page. Otherwise, it must also contain one word in the second list. After that the web page should also contain word in the third list. Actually these evaluations check not only lexical content but also CFT format information.

In this chapter, we discuss some methods for domain specific search. Based on these, we developed a domain specific search system. In chapter 5, we will present the system application and implementation.

Chapter 5
Implementation

In this chapter, we'll first introduce some software packages as the system construction tools. Then we discuss and compare different commercial search engines (CSE). Finally we talk about the system functions and some detail of implementation.

5.1 Survey of construction tools

The designing and construction of modern software is eased by the availability of high quality public domain system. For our applications, the following software packages were used: the HttpUnit package to be used in simulating a web browser to complete an URL connection; the Lucene will be introduced as a very good indexing tool for lexical analyzing; the HTMLParser package to be used in parsing and analyzing those downloaded Web pages; and the Package Regexp to be used in regular expression and the package HSQL Data Base to be used in data store.

5.1.1 HttpUnit

Automated testing is a great way to ensure that code being maintained works. HttpUnit is a suite of Java classes to test Web applications over HTTP. Coupled with Junit, HttpUnit is a powerful tool for creating test suites to ensure the end-to-end functionality of Web applications. One great aspect of HttpUnit is that it can test entire Web applications, not just single pages. HttpUnit makes it easy for a program to bypass the browser and access a web site from a program. Written in Java, HttpUnit emulates the relevant portions of browser behavior, including form submission, JavaScript, basic http authentication, cookies and automatic page redirection, and allows Java test code to examine returned pages either as text, an XML DOM, or containers of forms, tables, and links. Because it works closely with information and cookies, we can write tests to cover a whole session. For example, if our Web application includes a shopping cart, we could write a test to try logging in, selecting an item, placing it the shopping cart, and checking out. Since the tests are written in Java, there's no limit to how in-depth the tests can be. We introduce the main parts of HttpUnit in the following.

1) Junit

JUnit is a simple framework to write repeatable tests. JUnit is a program used to perform unit testing of virtually any Java software. JUnit testing is accomplished by writing test cases using Java, compiling these test cases and running the resultant classes with a JUnit Test Runner.

2) Making a request

Since HttpUnit can emulate an entire session and not just a single request, the system uses a class, called WebConversation, to manage the requests, handle the cookies, and resolve relative URLs. As we write more complicated tests or simulate a browser to connect with an URL covering a whole session, the WebConversation class will become more important.

3) Parsing a response

Once we've made a successful request to a Web server, it's time to parse the result of the request. HttpUnit makes use of the JTidy package, which is included in the HttpUnit distribution, to parse the resulting HTML into a Document Object Model (DOM) tree. For those who aren't familiar with DOM trees, they offer a uniform way to manipulate a document in a hierarchical data structure. JTidy provides a standardized way to manipulate the HTML result.

It should be noted that, however, the DOM tree that JTidy builds represents the structure that a document should have, not necessarily what it does have. That means that JTidy may add structural elements to the document tree that aren't in the HTML source but should be. That includes head and body markers, paragraph marks, font tags, and more. Iterating through a DOM tree and visualizing the output, can be helpful.

4) Shortcuts through the DOM

Navigating through the DOM can be difficult and time-consuming. Fortunately, HttpUnit includes tools to make quick work of dealing with some HTML elements. HttpUnit contains similar shortcuts for navigating the DOM tree for forms and tables.

5) Navigating links

HttpUnit offers much more than just parsing the result of a single connection. The real power comes from being able to make multiple requests through the WebConnection object. The easiest way to make multiple requests is by following HTML links.

6) Posting forms

By far, the most common use of HttpUnit is to test the results of filling out a form. We can fill out a form of interactive Web page to achieve the desired effect. The WebConversation object takes care of making sure that all the links go to the correct endpoint and that all cookies are honored through the session. Each request uses the same WebConversation object, which ensures that the cookie values stay the same, simulating an entire session. The output of the test will include all the headers from the server so we can see that the cookies are being sent only once by the server.

In a word, all these features make sure that HttpUnit meet not only the need of testing, but also the need of a Web page analyzing. The chosen version is HttpUnit 1.5.4, published at 21 Aug, 2003, sizing around 2.59M.

5.1.2 Lucene

Lucene is a Java library that adds text indexing and searching capabilities to an application. Lucene offers two main services: text indexing and text searching. These two activities are relatively independent of each other, although indexing naturally affects searching. The fundamental Lucene classes for indexing text are: *IndexWriter*, *Analyzer*, *Document*, and *Field*. IndexWriter is used to create a new index and to add Documents to an existing index. Before text is indexed, it is passed through an Analyzer. Analyzers are in charge of extracting indexable tokens out of text to be indexed, and eliminating the rest. Lucene comes with a few different Analyzer implementations. Some of them deal with skipping stop words (frequently-used words that don't help distinguish one document from the other, such as "a," "an," "the," "in," "on," etc.), some deal with converting all tokens to lowercase letters, so that searches are not case-sensitive, and some like PorterStemAnalyzer will perform Porter stemming (or "Porter stemmer") on its input, which is a process for removing the more common morphological and inflectional endings from words in English. For example, a group of words CONNECT, CONNECTED, CONNECTING, CONNECTION, CONNECTIONS will usually have similar meanings. After Porter stemming, the group of words are left CONNECT only. Its main function is to be part of a term normalization process that is usually done when setting up Information Retrieval systems. An index consists of a set of Documents, and each Document consists of one or more Fields. Each Field has a name and a value. Think of a Document as a row in a RDBMS, and Fields as columns in that row.

5.1.3 HTMLParser

HTML Parser is a library, written in Java, which allows to parse HTML (HTML 4.0 supported). It has been used by people on live projects. The architecture is flexible, allowing to extend it easily. It is a super-fast real-time parser for real-world HTML. What has attracted most users to HTMLParser has been its simplicity in design, speed and ability to handle streaming real-world html. The chosen version is version 1.3, published on 24 May, 2003, sizing around 1.8M bytes.

5.1.4 Regexp

Regexp is a 100% Pure Java Regular Expression package that was graciously donated to the Apache Software Foundation by Jonathan Locke. He originally wrote this software back in 1996 and it has stood up quite well to the test of time. It includes complete Javadoc documentation as well as a simple Applet for visual debugging and testing suite for compatibility. The chosen version is Regexp 1.3, sizing around 195K bytes. It is a sub-project of Jakarta.

5.1.5 Hsql Data Base

HSQL Database Engine (hsqldb) Project hosted at SourceForge. The hsqldb is a relational database engine written in Java, with a JDBC driver, supporting a rich subset of ANSI-92 SQL (BNF tree format). It offers a small (less than 160k), fast database engine which offers both in memory and disk based tables. Embedded and server modes are available. Additionally, it includes tools such as a minimal web server, in-memory query and management tools (can be run as applets) and a number of demonstration examples. The chosen version is HSQLDB version 1.7.1.

5.2 Choice of a commercial search engine (CSE.)

Normally, a web search application needs connect to URLs, downloads web pages, and analyzes them (Figure 5.2.1). But as a search research project, there is no need to implement the search application from scratch. The part of accessing HTML pages can be replace by selected CSE to download HTML pages in the Web.

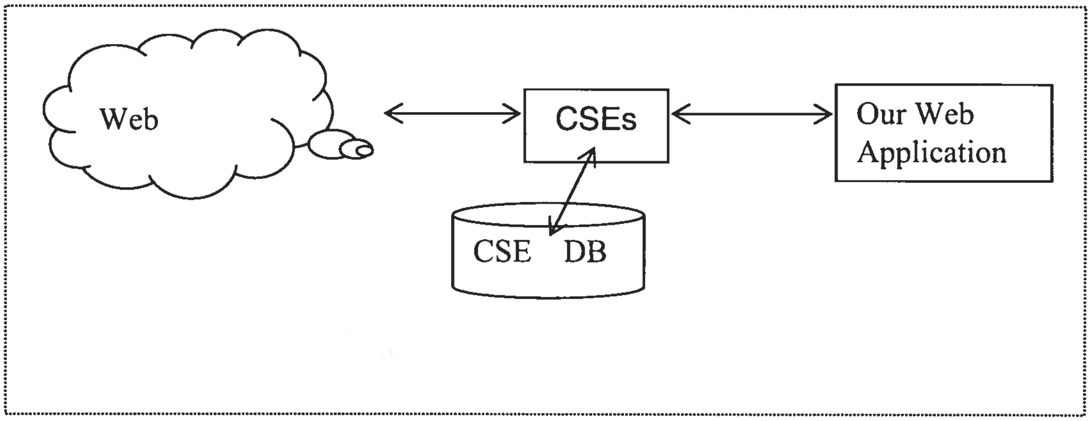


Figure 5.2.1 *Access HTML pages through CSE*

Currently, there are a lot of CSEs can be chosen. I tested 5 well known CSEs for both coverage and precision. These 5 CSEs are ALTAVISTA, GOOGLE, LYCOS, MSN, and YAHOO. In order to test the coverage of the five search engines, we first choose some typical query terms usually used in the "call for tender"(CFT) domain so as to check the searching results of different search engines. For out tests, we used the following term: "Ottawa tender" and "Ottawa RFQ" (request for quotation). The results are listed in table 5.2.1. The tables is ordered by the number of hits:

Order	CSE	Query Term	Coverage
1	GOOGLE	Ottawa tender Ottawa RFQ	57,000 2,490
2	YAHOO	Ottawa tender Ottawa RFQ	36,400 2,250
3	LYCOS	Ottawa tender Ottawa RFQ	26,466 552
3	ALTAVISTA	Ottawa tender Ottawa RFQ	20,196 496
3	MSN	Ottawa tender Ottawa RFQ	19,647 637

Table 5.2.1 Comparison of CSE (Coverage)

From *Table 5.2.1* we can see, for the different CSE, although the quantity of coverage varies with the query term, it roughly reflects the coverage of each CSE. Google and Yahoo cover respectively much more pages than the other three CSEs (ALTAVISTA LYCOS & MSN).

It is not enough to get many hits (coverage); a user wants useful and pertinent hits.

Order	CSE	Query Term	Coverage	Precision (at 20)
1	GOOGLE	Manitoba tender	33,700	35%
2	YAHOO	Manitoba tender	22,700	30%
3	ALTAVISTA	Manitoba tender	10,350	25%
4	MSN	Manitoba tender	10026	25%
5	LYCOS	Manitoba tender	13,396	10%

*Table 5.2.2 Comparison of CSE (*Coverage *& precision)*

To test this, we examined precision manually, examining the first 20 URLs returned by each search engine. We respectively input a query term “Manitoba tender” to each CSE (see the *Table 5.2.2.*). Google or Yahoo is also much higher than the other three. Based on these experiments data, the Google and Yahoo were chosen as the selecting CSE for the domain specific system.

5.3 **Introduction of the system**

The system was developed with Java 1.4.It amounts to 8179 lines of code and includes 29 classes in addition to the public library package. The appearance looks like the Figure 5.3.1. The overall view of menus are composed by four parts: “URL”(choose CSE) and “City” are set for the system start; “Depth” and “Option” are set for crawler; “Analysis” is experiment tools; Searching results can be checked in menu “CFT storage”, “CFT sites” , “CFT URL” & “Interactive CFT. We will first run the system then introduce the system functions. Before running the system, some parameters of the system need to be configured properly.

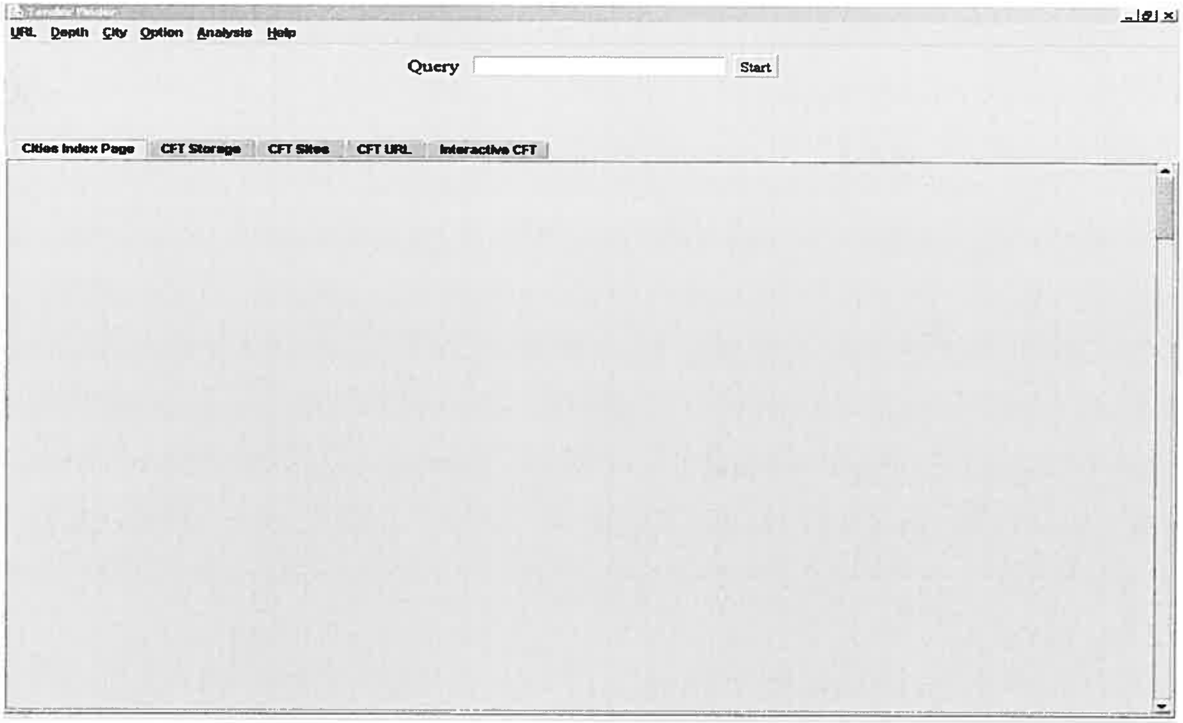


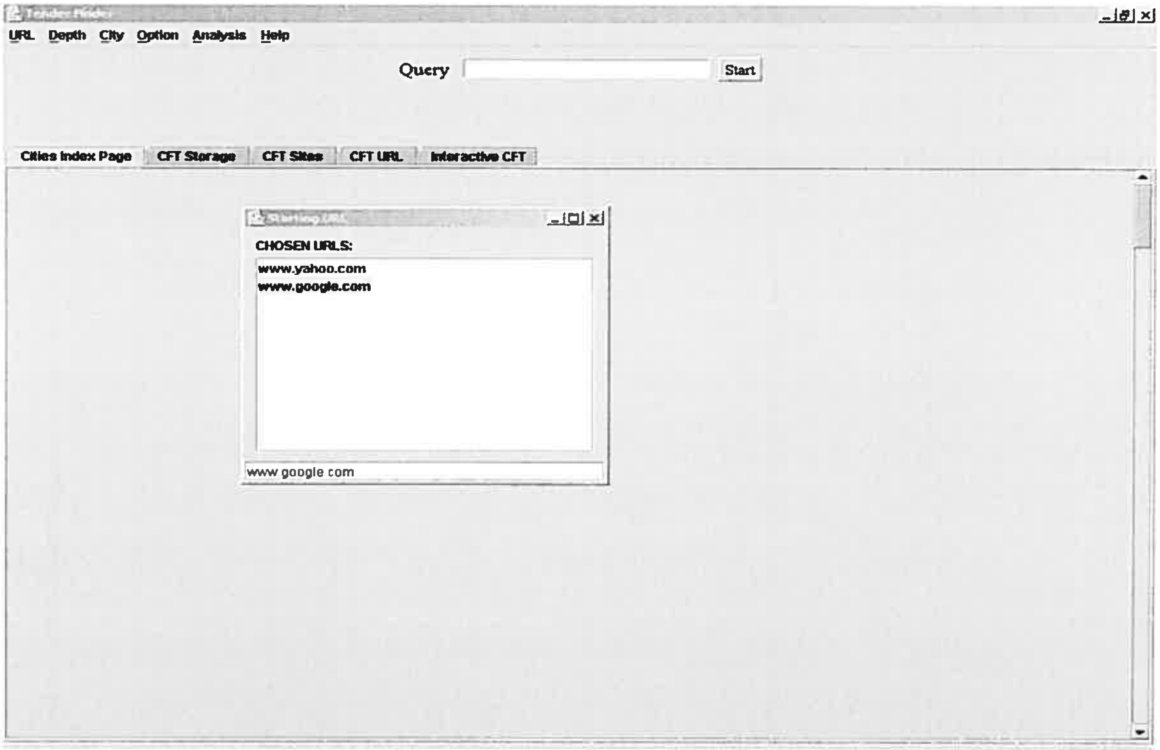
Figure 5.3.1 *Appearance of the system*

5.3.1 Parameters Configuration

Figure 5.3.1 shows various menus: URL, Depth, City, Option, Analysis and Help choice to be set. Let's examine them one by one.

1). URL

Both Google and Yahoo can be chosen as the CSE for initial request (Figure 5.3.2).



CHOSEN URLS:

www.yahoo.com
www.google.com

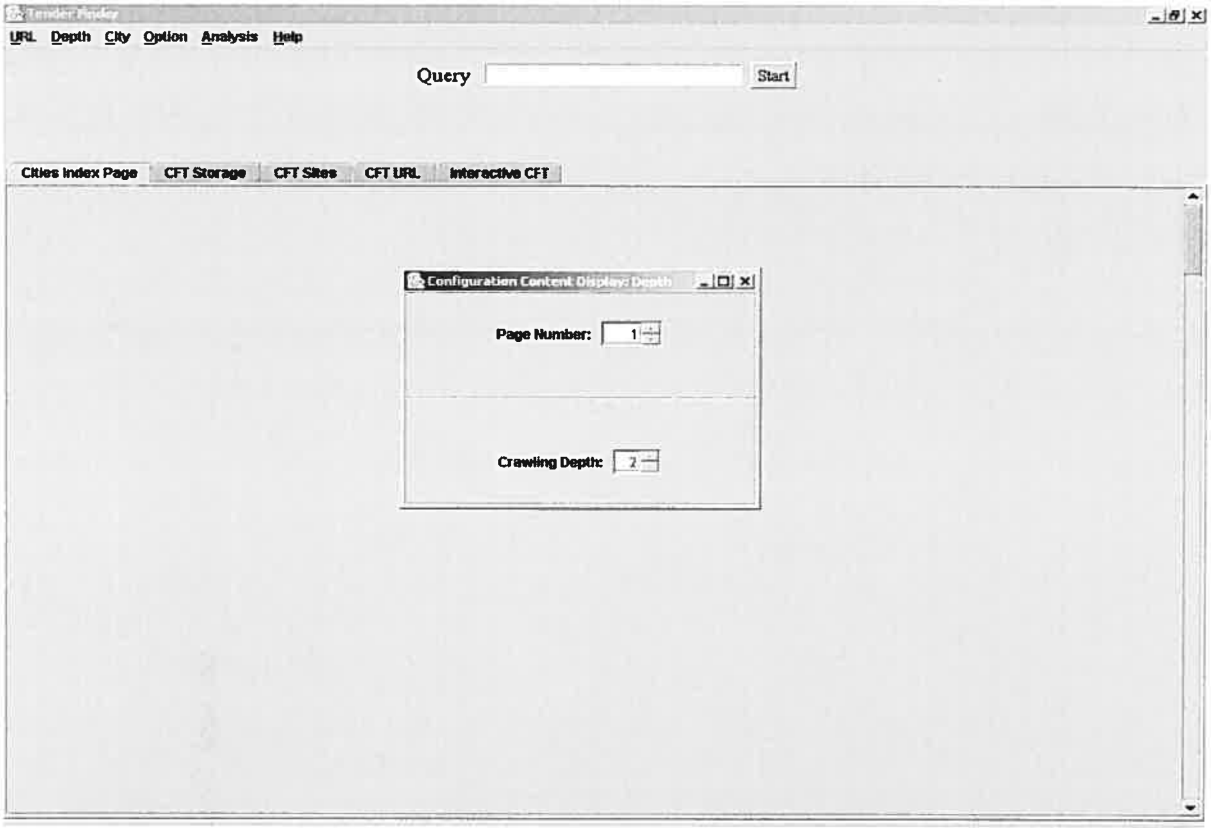
www google com

Figure 5.3.2 *URL configuration*

The setting process is, double clicking the menu URL, or pressing ALT+U, then appearing a pop-menu, double clicking the CSE in the window again.

2). Depth

In our system, the crawler will examine all interesting pages within specified distance (depth) from the initial links returned by the CSE. See the figure 5.3.3, under the Depth configuration there are two sub-items: "crawling depth" and "page number". Like the URL configuration, we can choose the "crawling depth" or "page number" respectively. The configuration of "crawling depth" decides layers a crawler further digging in. This parameter is the most time consuming factor. The more the "crawling depth" being chosen, the more time you need waiting, of course the more search results are provided.



Page Number: 1

Crawling Depth: 2

Figure 5.3.3 *Depth configuration*

The page number decides how much the CSE response pages will be chosen as seed pages.

3). City

In order to simplify the use and experimentation of our system in the Canadian context we have created a database of Canadian cities. Our system concentrates on one city at a time. To select a city, we can double click a "City" in city choosing pop-menu. Or we can select "No choice" to input in the query field later. When choosing a city, the chosen city will be combined with other query terms provides to the search system. See the figure 5.3.4 in the following page.

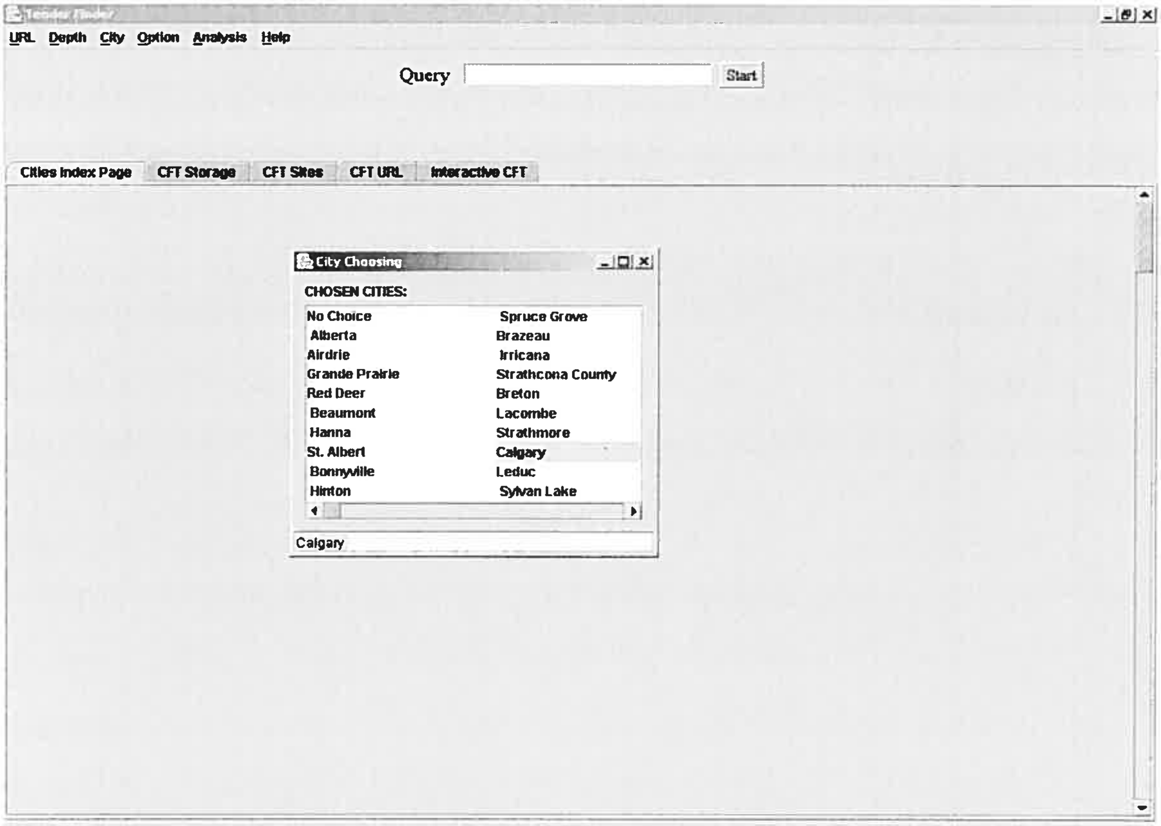


Figure 5.3.4 *City configuration*

4). Option

The configuration of "option" includes two sub-items: "Tender extension" and "Tender Filter". Tender extension aim to extend the query term so as to avoid tunneling and get more search results. You can input a word and add it to a list of "candidate words" and then press a button to forward the word to a list of "Extension Terms", and vice versa. See the following figure 5.3.5

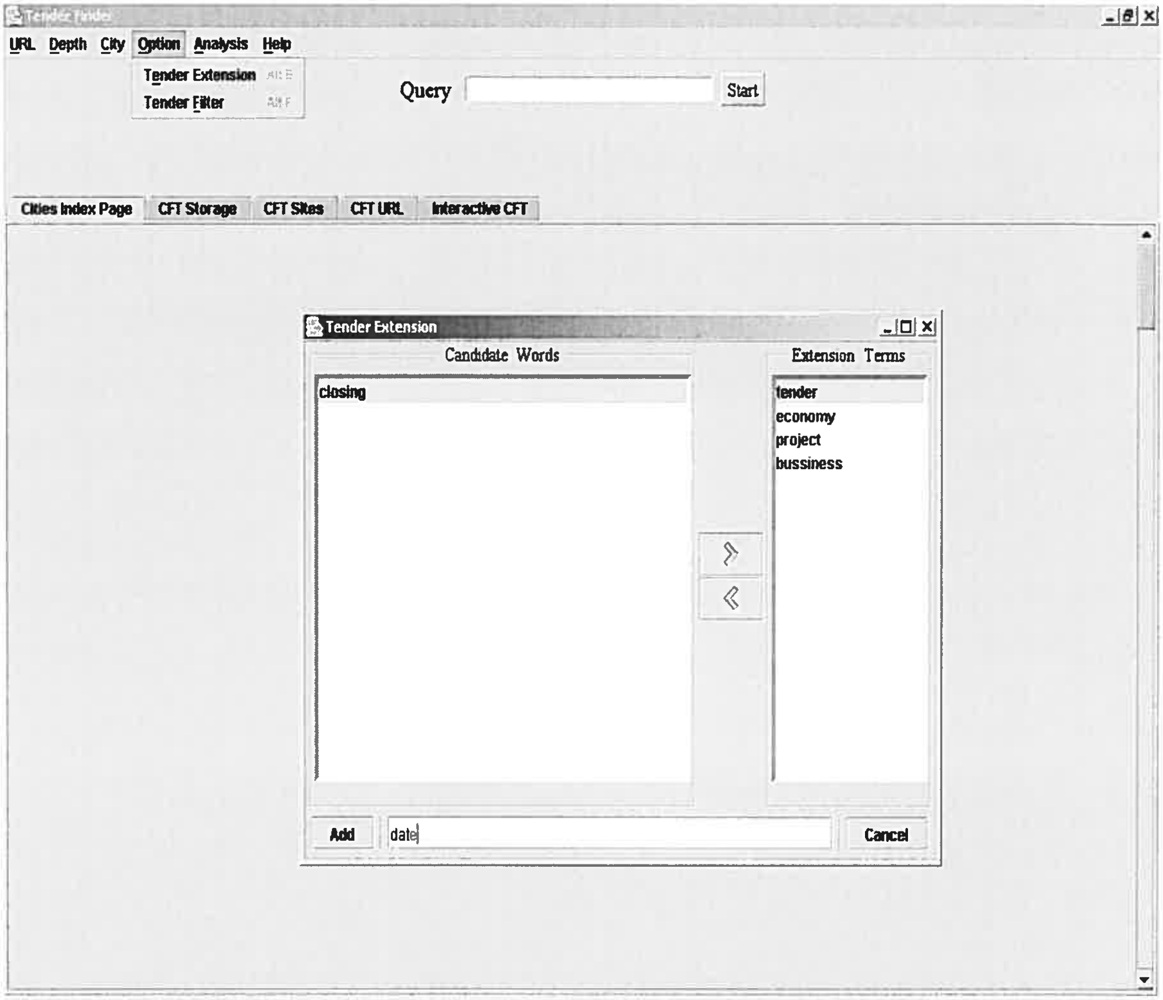
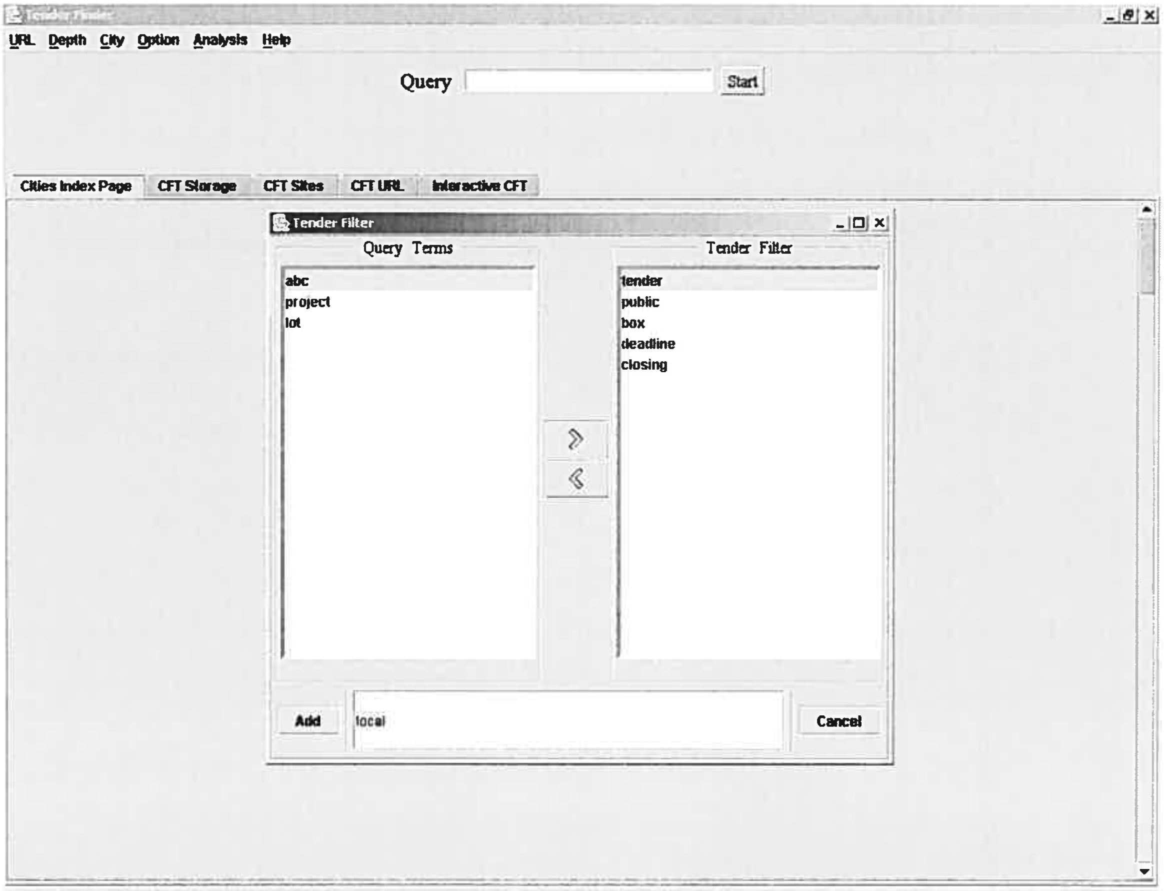


Figure 5.3.5 *Option configuration --- tender extension*

There are another option is "tender filter". You can add some word to a list of "tender filter" to filter the searching result. For example, you can add a word "project" to see how many frequencies appearing in some group of documents. This can be combined used with lexical analysis to check a word occurring frequency. See figure 5.3.6 in next page.

4). Analysis

This menu leads to the sub-system responsible for analyzing the HTML structure of a page to extract the portion that leads with the domain of interest. As shown in figure 5.3.6, analysis includes structure analysis and lexical analysis.



Query Terms: abc, project, lot

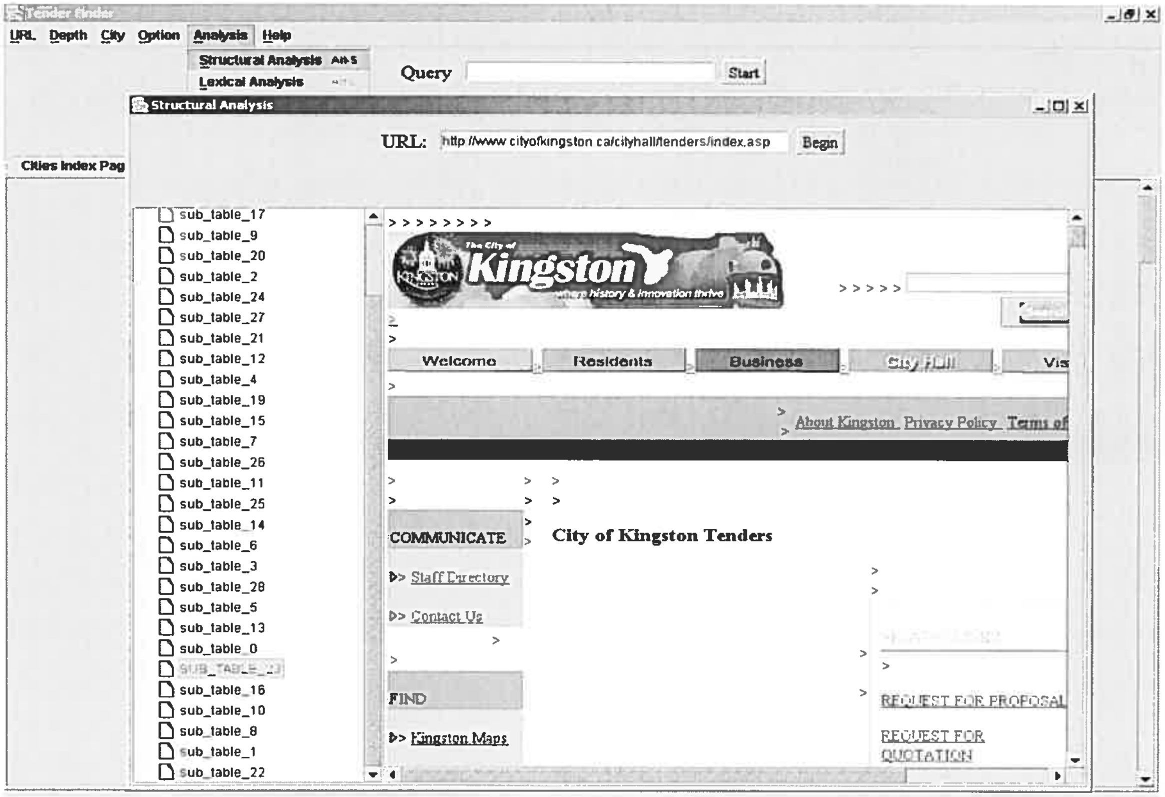
Tender Filter: tender, public, box, deadline, closing

local

Figure 5.3.6 Option *configuration --- tender filter*

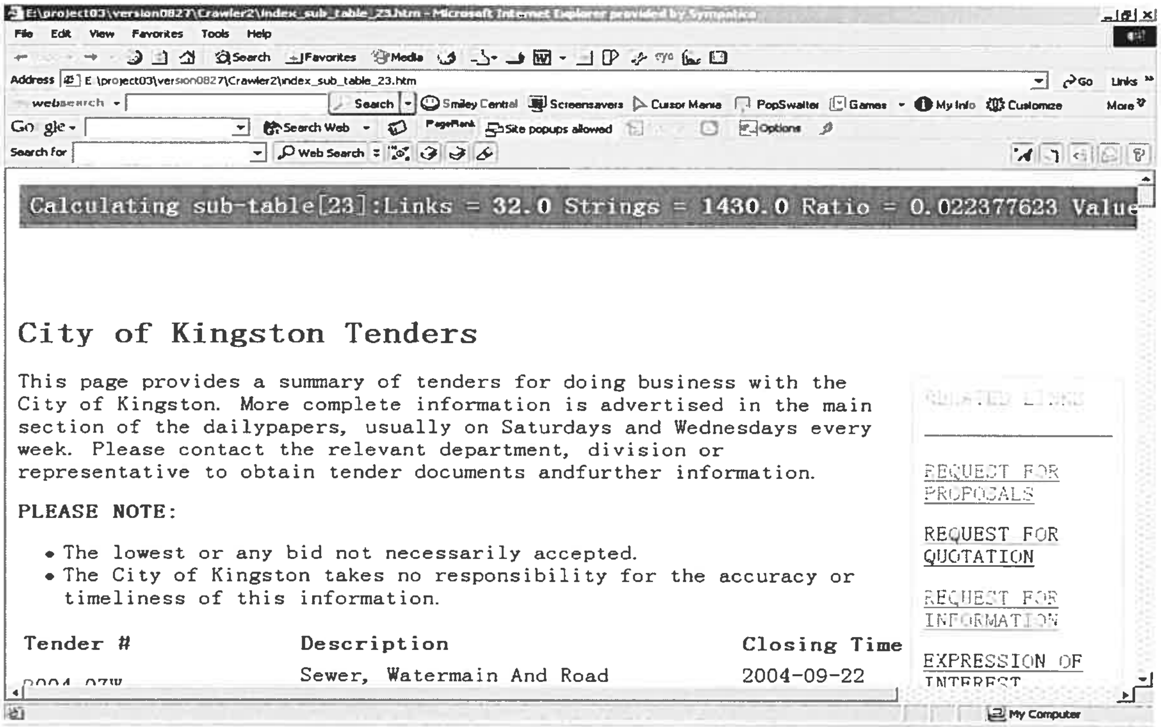
"Structure analysis" like figure 5.3.7 shown, you input an URL like: http://www.cityofkingston.ca/cityhall/tenders/index.asp and press the button "Begin". The system will explore that web page and analyze it. The right side of the pop window is content of that web page; the left side of the pop window is structural analyzing results, which marks sub-table with an order. One of these sub-tables is a selected one, which marks with capital letter like SUB_TABLE with a number 23. See figure 5.3.7 in next page. The SUB_TABLE_23 looks like figure 5.3.8 shown.

Necessity of analyzing logical structure for a CFT page based on the following considerations. The more exploring paths a crawler has the better searching coverage it has. On the other hand, while the crawler meets a qualified CFT page, it should bring the current exploring process to a temporary end, downloads this page, and starts another exploring process.



URL: http://www.cityofkingston.ca/cityhall/tenders/index.asp

Figure 5.3.7 *Analysis configuration --- structural analysis*


E:\project03\version0827\Crawler2\index_sub_table_23.htm - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Search Favorites Media
Address E:\project03\version0827\Crawler2\index_sub_table_23.htm
Go Links
websearch Search Smiley Central Screensavers Cursor Mania PopSwatter Games My Info Customize More
Google Search Web PageRank Site popups allowed Options
Search for Web Search
My Computer

```
Calculating sub-table[23]:Links = 32.0 Strings = 1430.0 Ratio = 0.022377623 Value
```

City of Kingston Tenders

This page provides a summary of tenders for doing business with the City of Kingston. More complete information is advertised in the main section of the dailypapers, usually on Saturdays and Wednesdays every week. Please contact the relevant department, division or representative to obtain tender documents andfurther information.

PLEASE NOTE:

- The lowest or any bid not necessarily accepted.
- The City of Kingston takes no responsibility for the accuracy or timeliness of this information.

Tender #	Description	Closing Time
2004-07W	Sewer, Watermain And Road	2004-09-22

RELATED LINKS

REQUEST FOR PROPOSALS

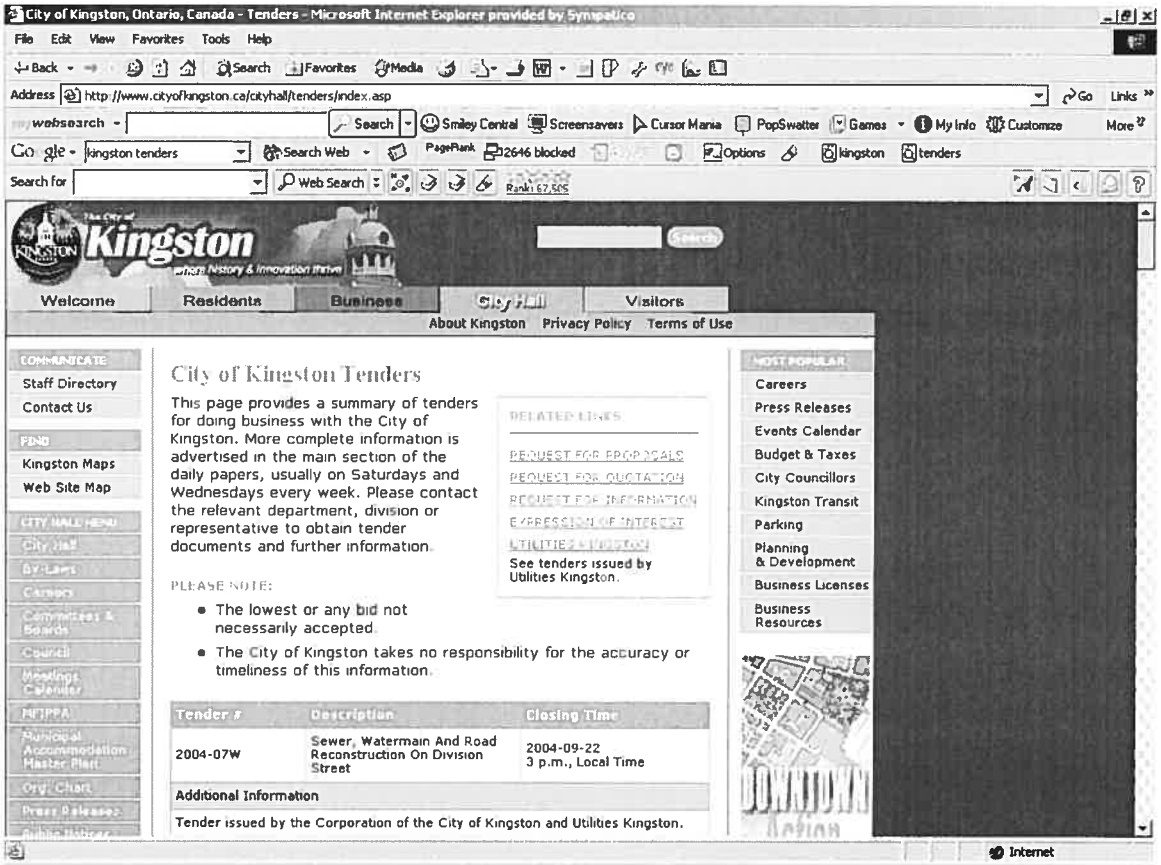
REQUEST FOR QUOTATION

REQUEST FOR INFORMATION

EXPRESSION OF INTEREST

Figure 5.3.8 *Selected sub-table content*

It is easy to process a CFT page which contains mainly text or mainly links. But for a mixed page like Figure 5.3.9, extracting a middle part of text content can improve accuracy of a lexical evaluation. Besides, such cutting process can also be used for analyzing a CFT page part by part.



City of Kingston Tenders

This page provides a summary of tenders for doing business with the City of Kingston. More complete information is advertised in the main section of the daily papers, usually on Saturdays and Wednesdays every week. Please contact the relevant department, division or representative to obtain tender documents and further information.

See tenders issued by Utilities Kingston.

PLEASE NOTE:

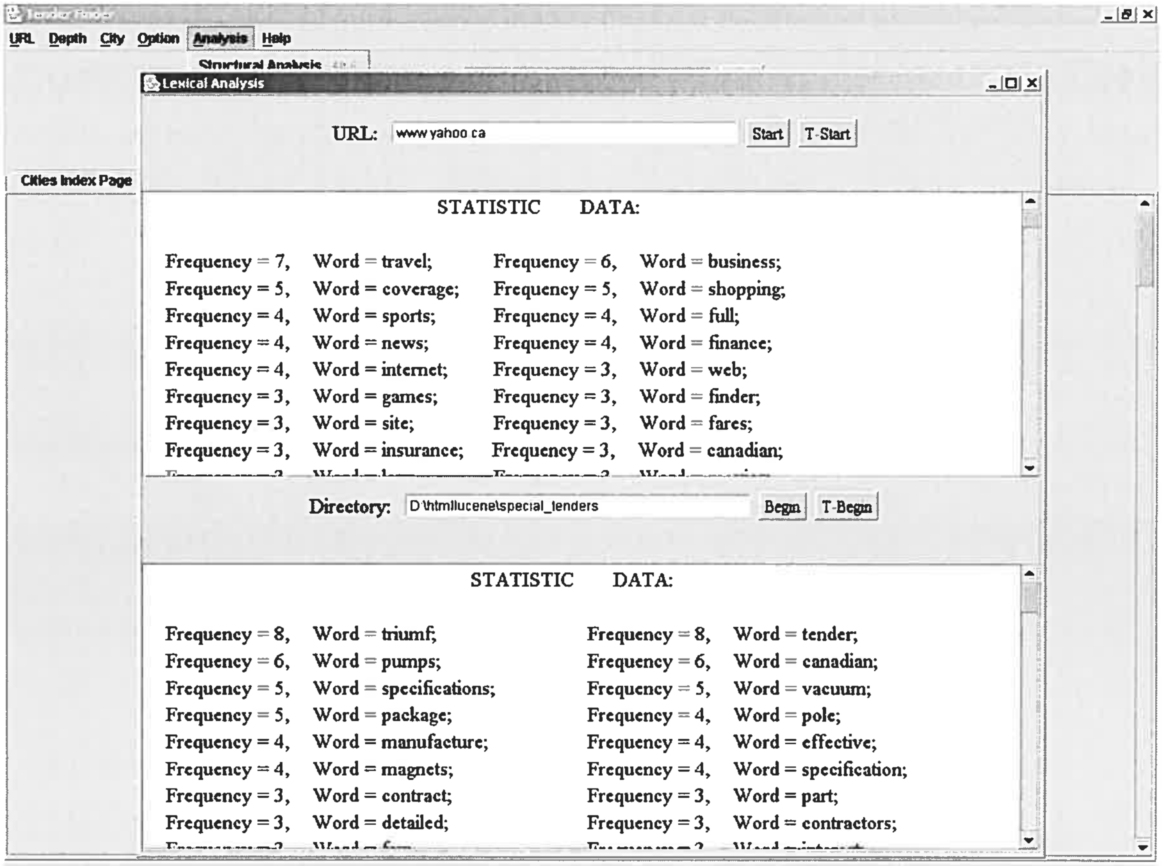
- The lowest or any bid not necessarily accepted.
- The City of Kingston takes no responsibility for the accuracy or timeliness of this information.

Tender #	Description	Closing Time
2004-07W	Sewer, Watermain And Road Reconstruction On Division Street	2004-09-22 3 p.m., Local Time
Additional Information		
Tender issued by the Corporation of the City of Kingston and Utilities Kingston.		

Figure 5.3.9 *Mixed page*

The detail of decomposing mixed page to several sub-tables has been explained in chapter 4.

Figure 5.3.10 in next page shows the lexical analysis process.



URL: www yahoo ca

```
STATISTIC      DATA:

Frequency = 7,   Word = travel;       Frequency = 6,   Word = business;
Frequency = 5,   Word = coverage;     Frequency = 5,   Word = shopping;
Frequency = 4,   Word = sports;       Frequency = 4,   Word = full;
Frequency = 4,   Word = news;         Frequency = 4,   Word = finance;
Frequency = 4,   Word = internet;     Frequency = 3,   Word = web;
Frequency = 3,   Word = games;        Frequency = 3,   Word = finder;
Frequency = 3,   Word = site;         Frequency = 3,   Word = fares;
Frequency = 3,   Word = insurance;    Frequency = 3,   Word = canadian;
```

Directory: D \htmllucene\special_tenders

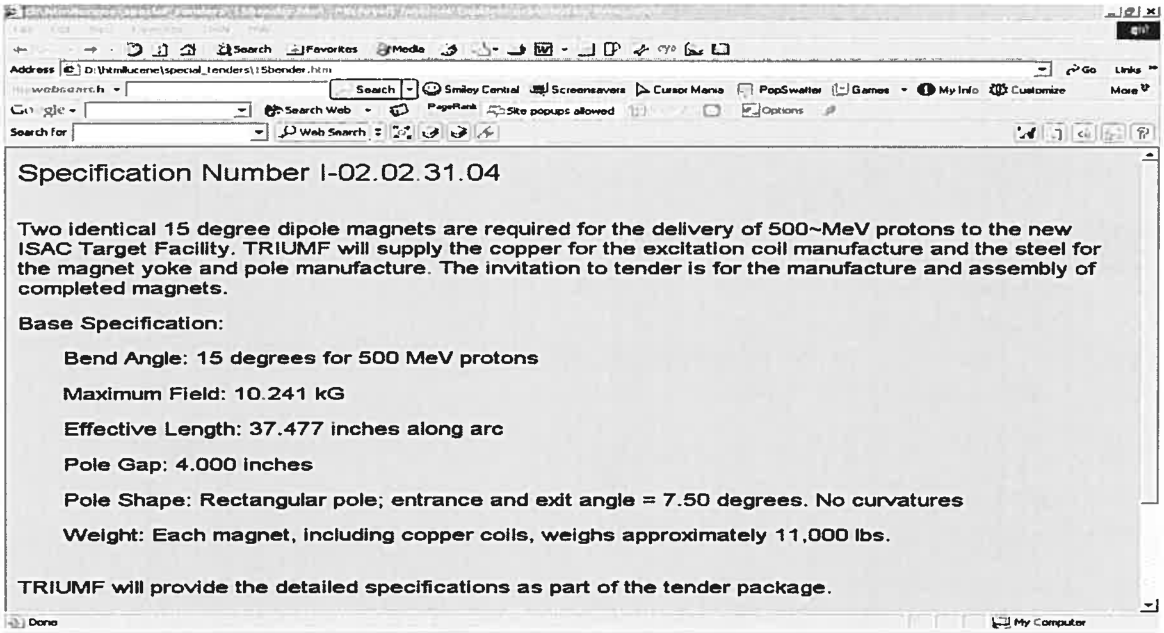
```
STATISTIC      DATA:

Frequency = 8,   Word = triumf;             Frequency = 8,   Word = tender;
Frequency = 6,   Word = pumps;              Frequency = 6,   Word = canadian;
Frequency = 5,   Word = specifications;     Frequency = 5,   Word = vacuum;
Frequency = 5,   Word = package;            Frequency = 4,   Word = pole;
Frequency = 4,   Word = manufacture;        Frequency = 4,   Word = effective;
Frequency = 4,   Word = magnets;            Frequency = 4,   Word = specification;
Frequency = 3,   Word = contract;           Frequency = 3,   Word = part;
Frequency = 3,   Word = detailed;           Frequency = 3,   Word = contractors;
```

Figure 5.3.10 *Analysis configuration --- lexical analysis*

The lexical analysis is a statistic tool for analyzing various web pages. In the above window, the top window is for analyzing a web pages or a set of pages, the bottom window is for analyzing a saved example web file or a local directory which contains many web files. After inputting files in a URL or a Directory field, pressing the button “Start” or “Begin” will produces results as the above figure showing. These results count all the words in documents and their frequency of occurrence.

The directory d:\htmllucene\special_tenders includes three downloaded web pages of Vancouver area. Their contents are shown in the next page, respective figure 5.3.11.a, figure 5.3.11.b and figure 5.3.11.c.



Specification Number I-02.02.31.04

Two identical 15 degree dipole magnets are required for the delivery of 500~MeV protons to the new ISAC Target Facility. TRIUMF will supply the copper for the excitation coil manufacture and the steel for the magnet yoke and pole manufacture. The invitation to tender is for the manufacture and assembly of completed magnets.

Base Specification:

Bend Angle: 15 degrees for 500 MeV protons

Maximum Field: 10.241 kG

Effective Length: 37.477 inches along arc

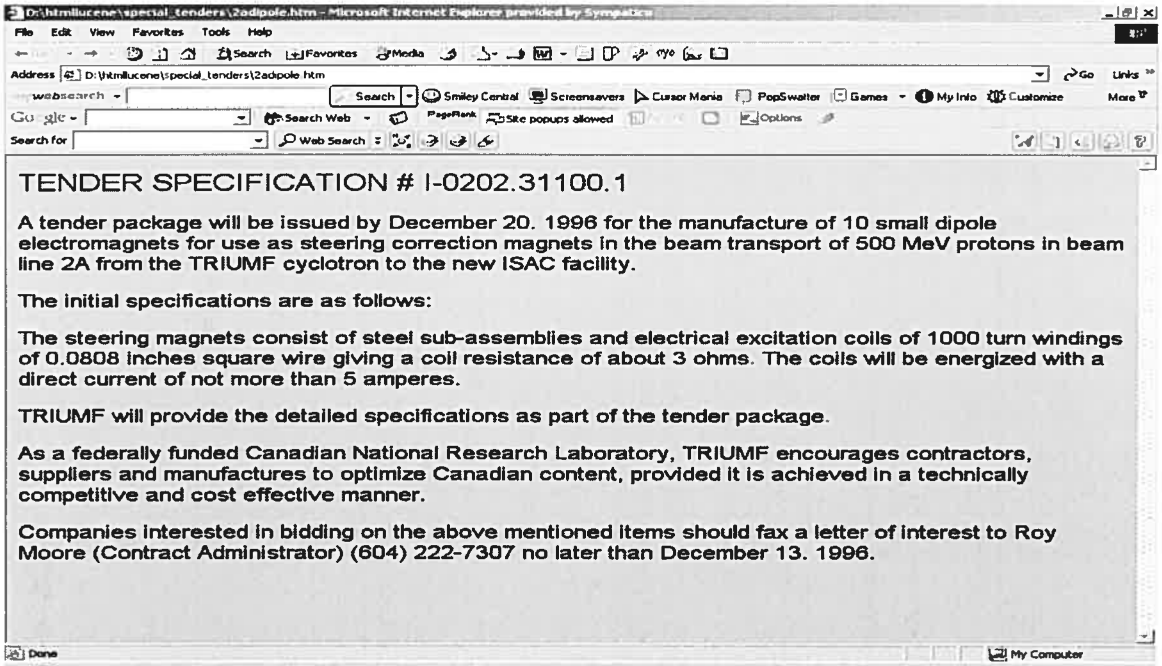
Pole Gap: 4.000 inches

Pole Shape: Rectangular pole; entrance and exit angle = 7.50 degrees. No curvatures

Weight: Each magnet, including copper coils, weighs approximately 11,000 lbs.

TRIUMF will provide the detailed specifications as part of the tender package.

Figure 5.3.11.a *Saved tender sample of Vancouver*


D:\htmllucene\special_tenders\2adipole.htm - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Address D:\htmllucene\special_tenders\2adipole.htm
Done
My Computer

TENDER SPECIFICATION # I-0202.31100.1

A tender package will be issued by December 20. 1996 for the manufacture of 10 small dipole electromagnets for use as steering correction magnets in the beam transport of 500 MeV protons in beam line 2A from the TRIUMF cyclotron to the new ISAC facility.

The initial specifications are as follows:

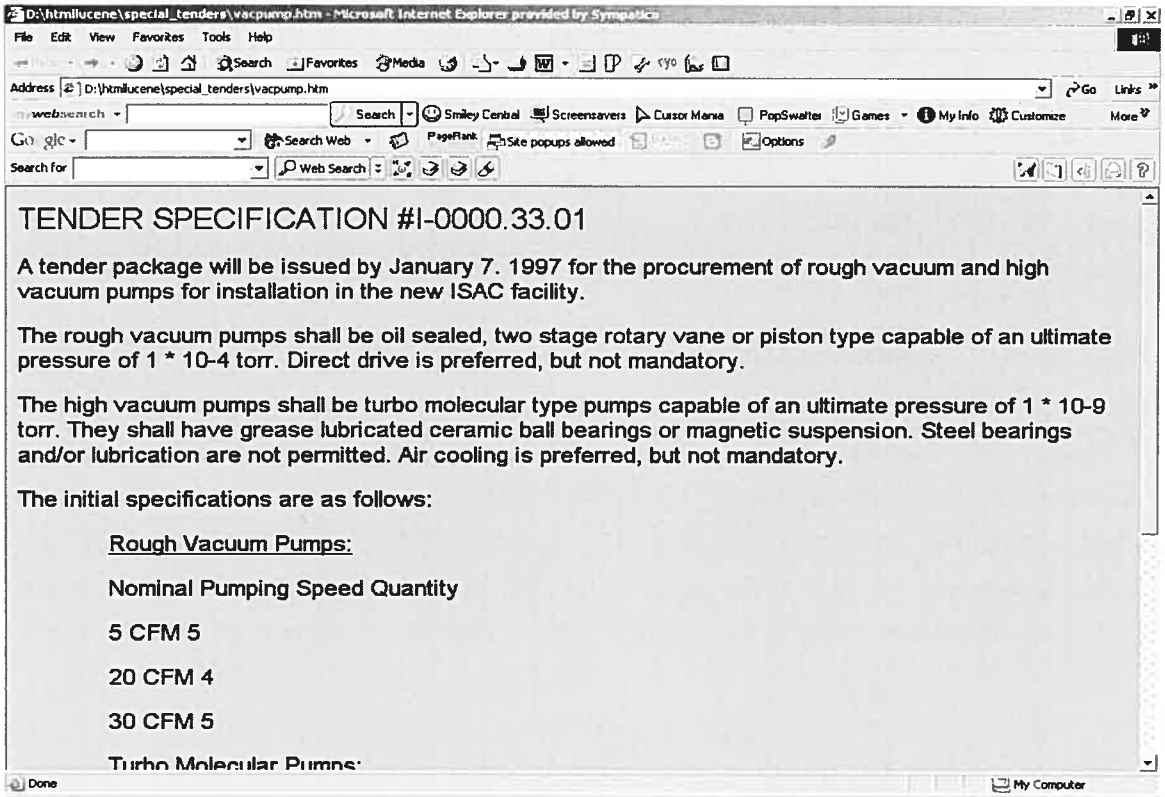
The steering magnets consist of steel sub-assemblies and electrical excitation coils of 1000 turn windings of 0.0808 inches square wire giving a coil resistance of about 3 ohms. The coils will be energized with a direct current of not more than 5 amperes.

TRIUMF will provide the detailed specifications as part of the tender package.

As a federally funded Canadian National Research Laboratory, TRIUMF encourages contractors, suppliers and manufactures to optimize Canadian content, provided it is achieved in a technically competitive and cost effective manner.

Companies interested in bidding on the above mentioned items should fax a letter of interest to Roy Moore (Contract Administrator) (604) 222-7307 no later than December 13. 1996.

Figure 5.3.11.b *Saved tender sample of Vancouver*


D:\htmllucene\special_tenders\vacpump.htm - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Search Favorites Media
Address D:\htmllucene\special_tenders\vacpump.htm
Go Links
websearch Search Smiley Central Screensavers Cursor Mania PopSwatter Games My Info Customize More
Google Search Web PageRank Site popups allowed Options
Search for Web Search
Done
My Computer

TENDER SPECIFICATION #I-0000.33.01

A tender package will be issued by January 7. 1997 for the procurement of rough vacuum and high vacuum pumps for installation in the new ISAC facility.

The rough vacuum pumps shall be oil sealed, two stage rotary vane or piston type capable of an ultimate pressure of 1 * 10-4 torr. Direct drive is preferred, but not mandatory.

The high vacuum pumps shall be turbo molecular type pumps capable of an ultimate pressure of 1 * 10-9 torr. They shall have grease lubricated ceramic ball bearings or magnetic suspension. Steel bearings and/or lubrication are not permitted. Air cooling is preferred, but not mandatory.

The initial specifications are as follows:

Rough Vacuum Pumps:

Nominal Pumping Speed Quantity

5 CFM 5

20 CFM 4

30 CFM 5

Turbo Molecular Pumps:

Figure 5.3.11.c Saved tender sample of Vancouver

The lexical analysis also provides “T-Start” and “T-Begin” buttons which trigger analysis according to pre-selected key words pertinent, in our case, to tenders specifically, the system will output the frequency in the document of keywords entered previously under the “tender filter” option. See the following figure 5.3.12a and figure 5.3.12b. In the figure 5.3.12a, we choose “lot”, “tender”, “project” and “deadline” as our filter words. In the figure 5.3.12b, the result shows that only “lot”, “tender” and “project” occurring frequencies. Same result when input a directory and pressing button “T-Begin”.

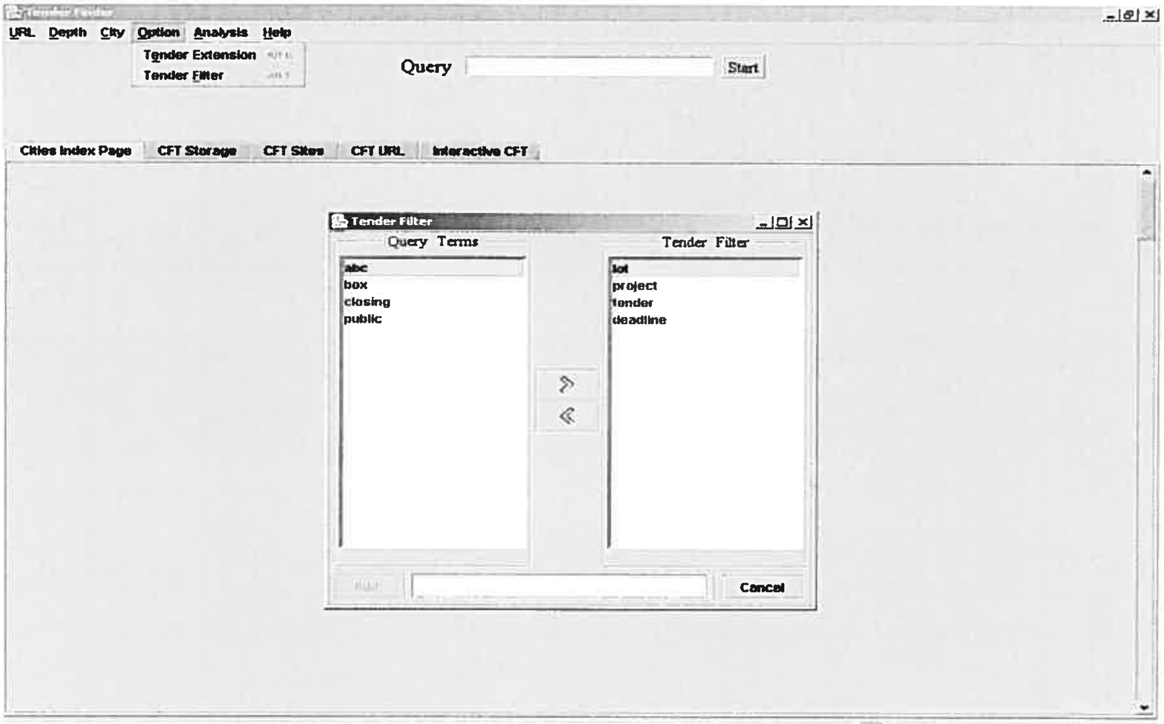
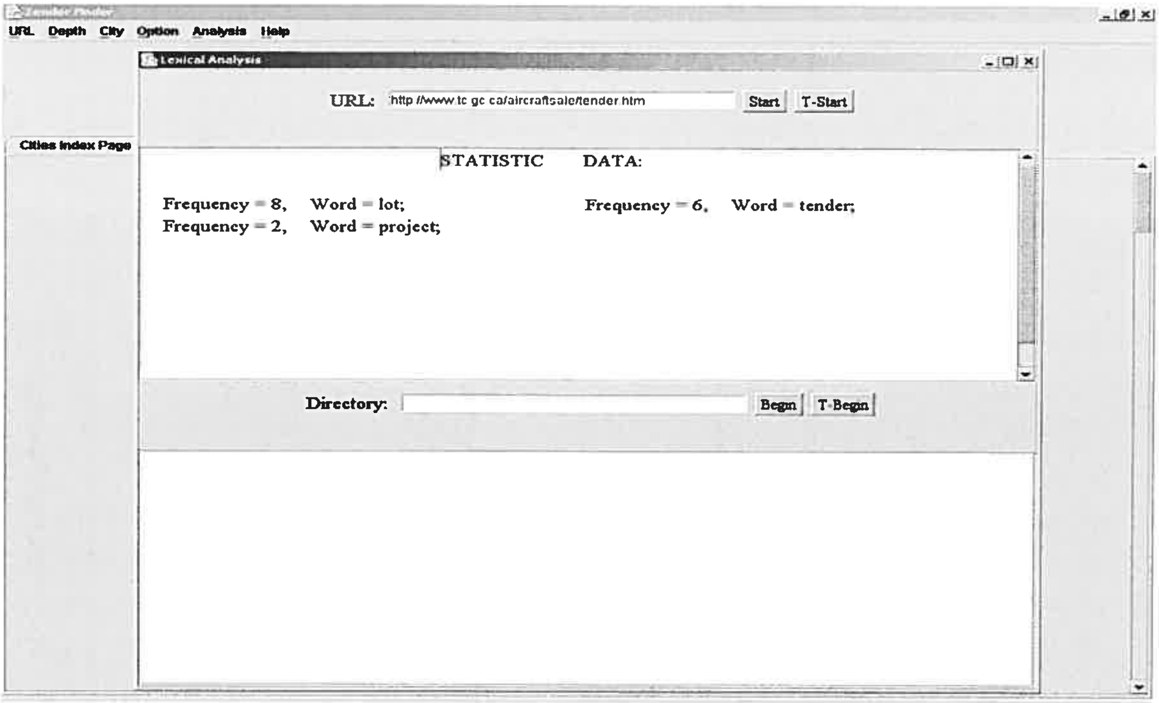


Figure 5.3.12a *Choice of tender filter*



```
                              STATISTIC   DATA:

Frequency = 8,   Word = lot;                 Frequency = 6,   Word = tender;
Frequency = 2,   Word = project;
```

Figure 5.3.12b *T-Start result*

5). Help

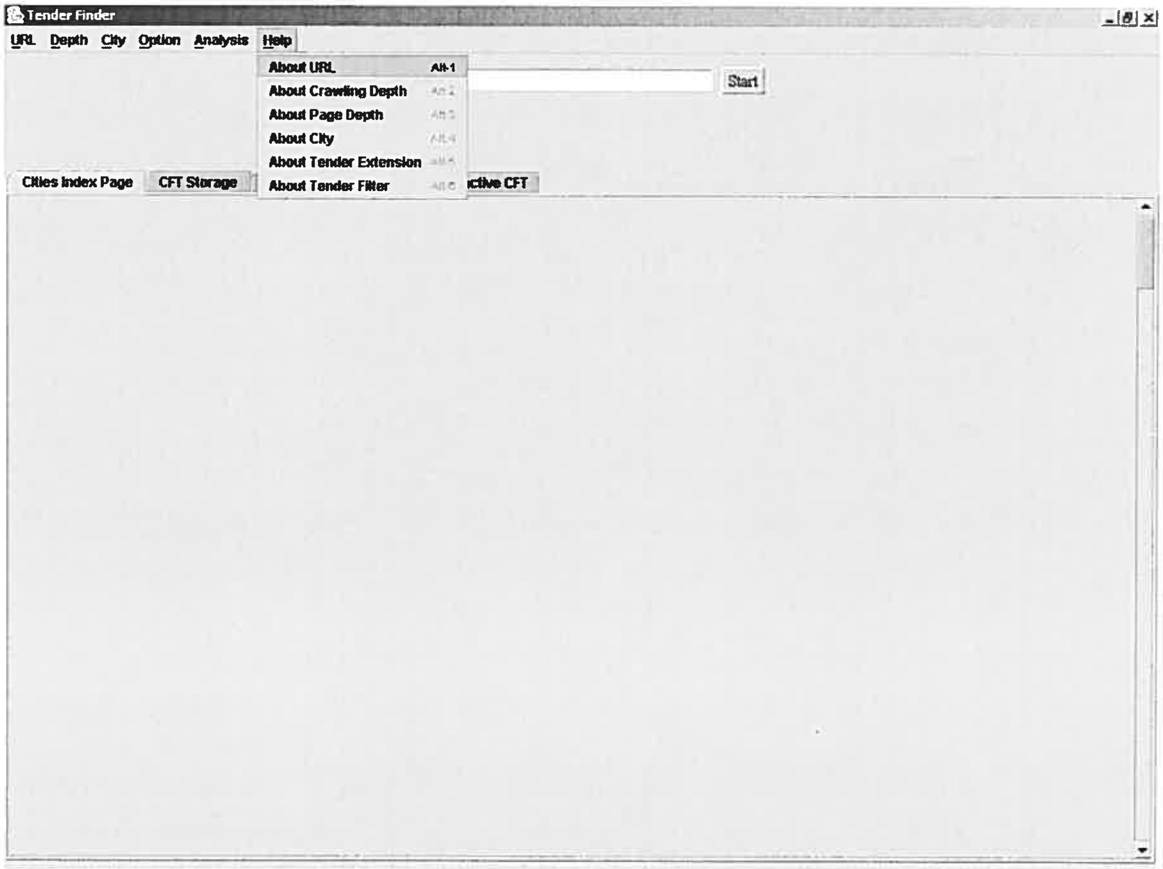


Figure 5.3.13 *Help menu*

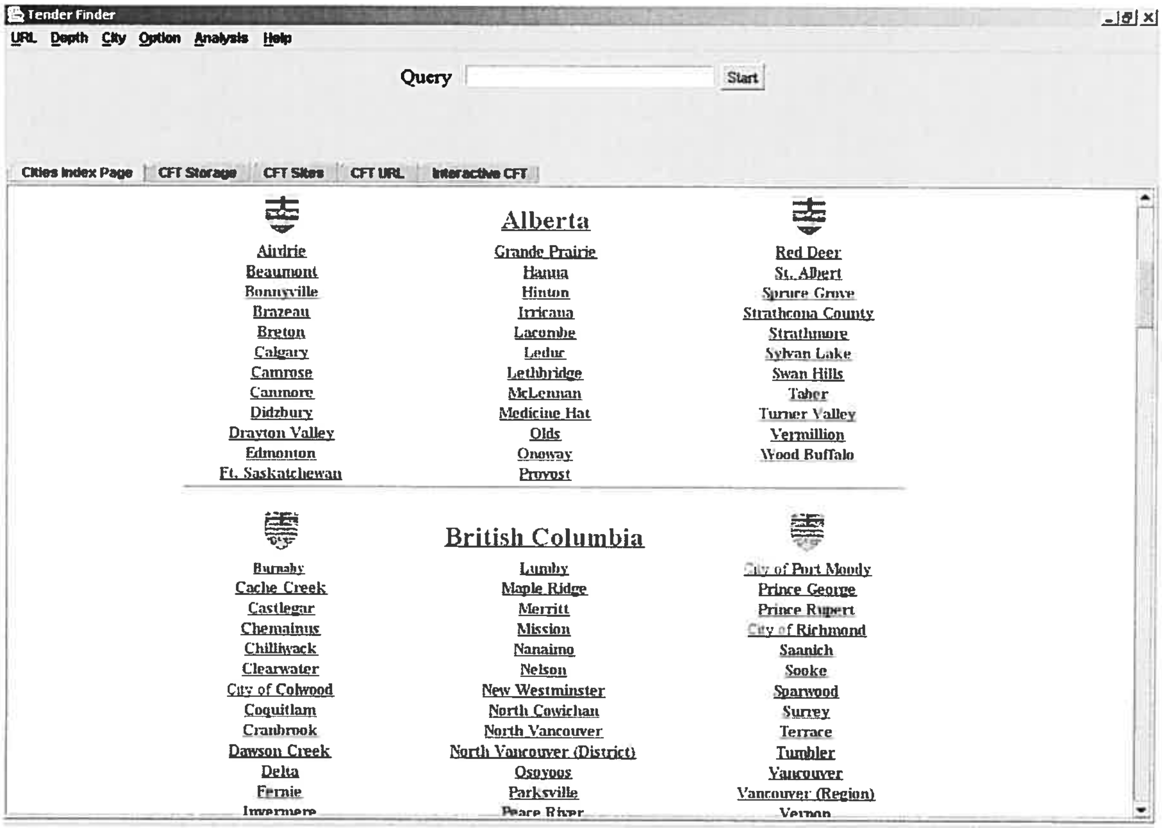
The "help" provides explanations for all the parameter configurations. It helps user to configure properly.

5.3.1 Running the system

We start a session by typing some keywords into the query field and pressing "start". Typically query would be a city with a term "tender" (you can input others if you choose "No choice" in the city pop-menu). This causes initial seed URLs to be obtained from Google or Yahoo then our crawler explores all interesting links to the specified depth limit and stores its results in the local database. When it has finished its exploration, the results can be examined via the TABBED sub-menus of the bottom window. We now illustrate the functionality of our system by going through a typical session The system includes 5 tabbed sub-menus which represent 5 different functions: Cities index page, CFT storage, CFT sites, CFT URL and interactive CFT. We'll introduce them one by one.

1). Cities index page

The index page is a site which can help us to navigate further to each province or city's home page (see figure 5.3.1a, figure 5.3.1b).



Alberta

Airdrie
Beaumont
Bonnyville
Brazeau
Breton
Calgary
Camrose
Canmore
Didzbury
Drayton Valley
Edmonton
Ft. Saskatchewan

Grande Prairie
Hanna
Hinton
Irricana
Lacombe
Ledur
Lethbridge
McLennan
Medicine Hat
Olds
Onoway
Provost

Red Deer
St. Albert
Spruce Grove
Strathcona County
Strathmore
Sylvan Lake
Swan Hills
Taber
Turner Valley
Vermillion
Wood Buffalo

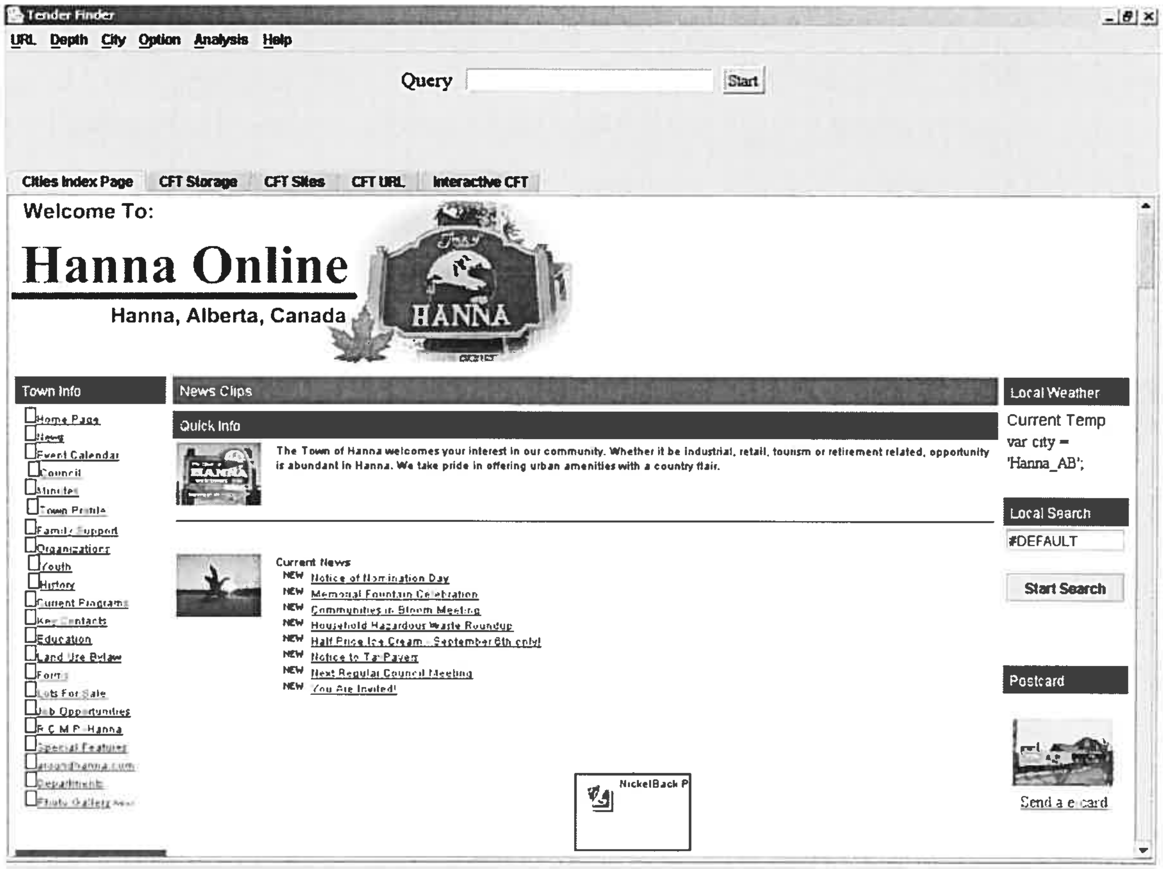
British Columbia

Burnaby
Cache Creek
Castlegar
Chemainus
Chilliwack
Clearwater
City of Colwood
Coquitlam
Cranbrook
Dawson Creek
Delta
Fernie
Invermere

Lumby
Maple Ridge
Merritt
Mission
Nanaimo
Nelson
New Westminster
North Cowichan
North Vancouver
North Vancouver (District)
Osoyoos
Parksville
Peace River

City of Port Moody
Prince George
Prince Rupert
City of Richmond
Saanich
Sooke
Sparwood
Surrey
Terrace
Tumbler
Vancouver
Vancouver (Region)
Vernon

Figure 5.3.1a *City Index Page*



The Town of Hanna welcomes your interest in our community. Whether it be industrial, retail, tourism or retirement related, opportunity is abundant in Hanna. We take pride in offering urban amenities with a country flair.

Current Temp
var city = 'Hanna_AB';

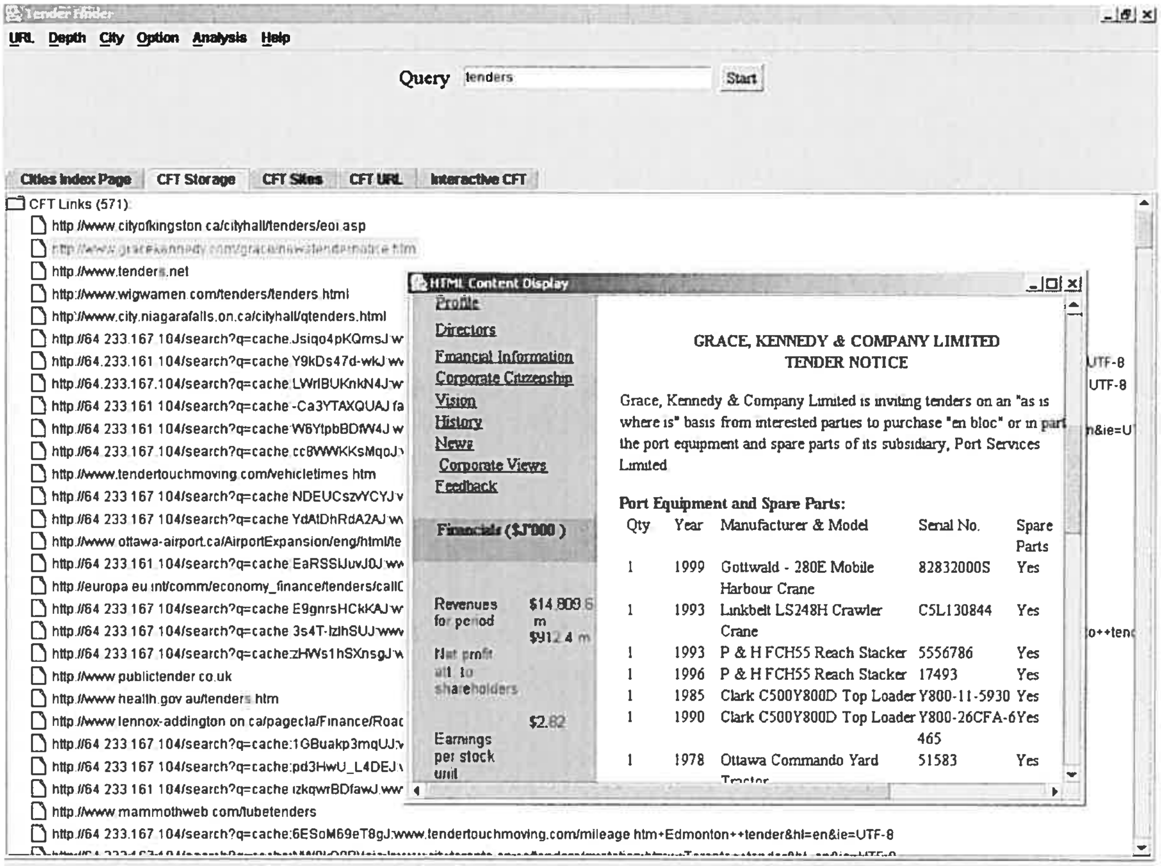
#DEFAULT

Figure 5.3.1b *City Index Page*

This is an auxiliary function for user to navigate to Canadian cities so as to access home page of each city. Figure 5.3.1b shows the result of selecting Hanna, Alberta.

2). CFT Storage

Clicking on the CFT storage tab shows the contents of the database (see Figure 5.3.2a) and clicking on an entry opens a further window showing the contents of the selected file. The crawler stores all the tender related web pages to a database according to content of pages or a URL string of a hyperlink. If a web page is a CFT page or "tender" is contained in the URL string, then the URL will be store in a database. See the Figure 5.3.2a in next page.



CFT Links (571)
http://www.cityofkingston ca/cityhall/tenders/eoi asp
http://www.gracekennedy.com/grace/newstendernotice.htm
http://www.tenders.net
http://www.wigwamen.com/tenders/tenders html
http://www.city.niagarafalls.on.ca/cityhall/qtenders.html
http://64 233.167 104/search?q=cache.Jsiqo4pKQmsJ:w
http://64.233.161 104/search?q=cache Y9kDs47d-wkJ ww
http://64.233.167.104/search?q=cache LWrlBUKnkN4J:w
http://64 233 161 104/search?q=cache -Ca3YTAXQUAJ fa
http://64 233 161 104/search?q=cache W6YtpbBDfW4J w
http://64 233.167 104/search?q=cache cc8WWKKsMqoJ:
http://www.tendertouchmoving.com/vehicletimes htm
http://64 233 167 104/search?q=cache NDEUCszvYCYJ:v
http://64 233 167 104/search?q=cache YdAtDhRdA2AJ ww
http://www ottawa-airport.ca/AirportExpansion/eng/html/te
http://64 233 161 104/search?q=cache EaRSSIJuvJ0J ww
http://europa eu int/comm/economy_finance/tenders/callC
http://64 233 167 104/search?q=cache E9gnrsHCkKAJ w
http://64 233 167 104/search?q=cache 3s4T-lzlhSUJ www
http://64 233 167 104/search?q=cache:zHWs1hSXnsgJ w
http://www publictender co.uk
http://www health.gov au/tenders htm
http://www lennox-addington on ca/pagecla/Finance/Roac
http://64 233 167 104/search?q=cache:1GBuakp3mqUJ:v
http://64 233 167 104/search?q=cache:pd3HwU_L4DEJ:
http://64 233 161 104/search?q=cache izkqwrBDfawJ ww
http://www mammothweb com/lubetenders
http://64 233.167 104/search?q=cache:6ESoM69eT8gJ:www.tendertouchmoving.com/mileage htm+Edmonton++tender&hl=en&ie=UTF-8

UTF-8
UTF-8
n&ie=U
o++tenc

Financials ($J'000)

Revenues for period $14 809.6 m
$912.4 m

Net profit att. to shareholders

$2.82

Earnings per stock unit

GRACE, KENNEDY & COMPANY LIMITED
TENDER NOTICE

Grace, Kennedy & Company Limited is inviting tenders on an "as is where is" basis from interested parties to purchase "en bloc" or in part the port equipment and spare parts of its subsidiary, Port Services Limited

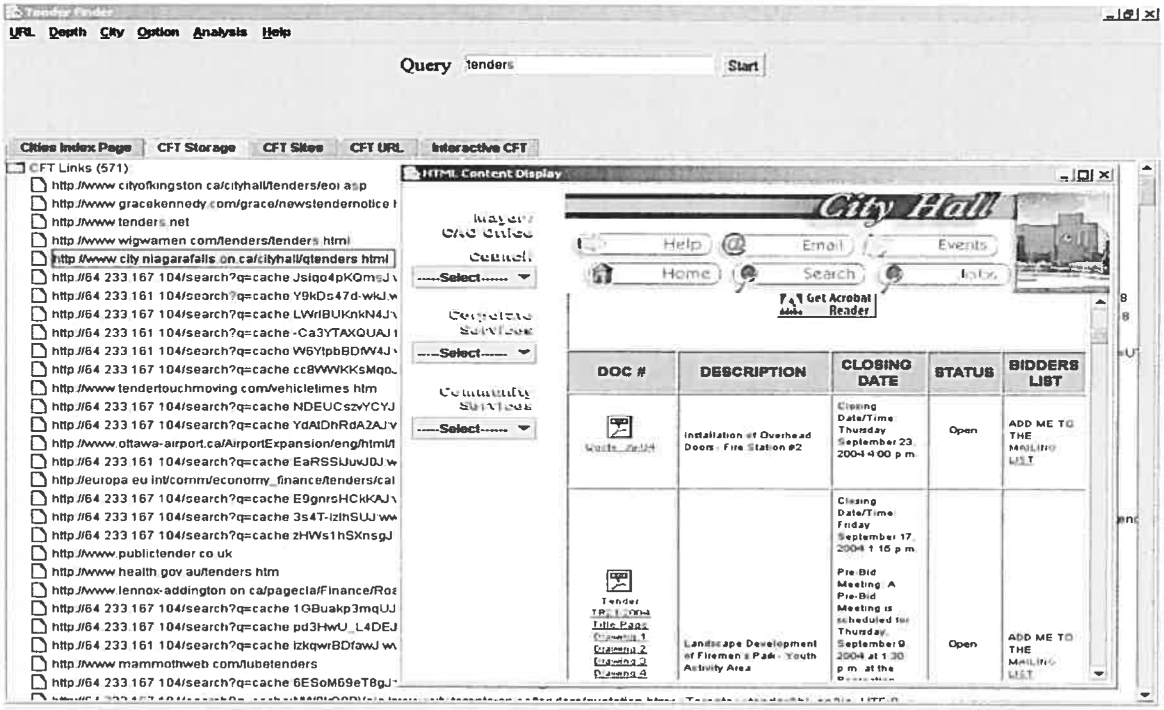
Port Equipment and Spare Parts:

Qty	Year	Manufacturer & Model	Serial No.	Spare Parts
1	1999	Gottwald - 280E Mobile Harbour Crane	82832000S	Yes
1	1993	Linkbelt LS248H Crawler Crane	C5L130844	Yes
1	1993	P & H FCH55 Reach Stacker	5556786	Yes
1	1996	P & H FCH55 Reach Stacker	17493	Yes
1	1985	Clark C500Y800D Top Loader	Y800-11-5930	Yes
1	1990	Clark C500Y800D Top Loader	Y800-26CFA-6465	Yes
1	1978	Ottawa Commando Yard Tractor	51583	Yes

Figure 5.3.2a *CFT Storage --- a CFT pertinent page*

The clicked URL string in the figure 5.3.2a is: http://www.gracekennedy.com/grace/newstendernotice.htm. It is a CFT page.

Figure 5.3.2b, Figure 5.3.2c and Figure 5.3.2d are other three random examples within the CFT storages. The first two are highly pertinent to CFT pages. The third is not. The check of most URLs within these results shows that those pointed pages are highly pertinent to CFT content. Thus, a URL string can be used in a fast content check before a pointed web page being downloaded.



Query tenders

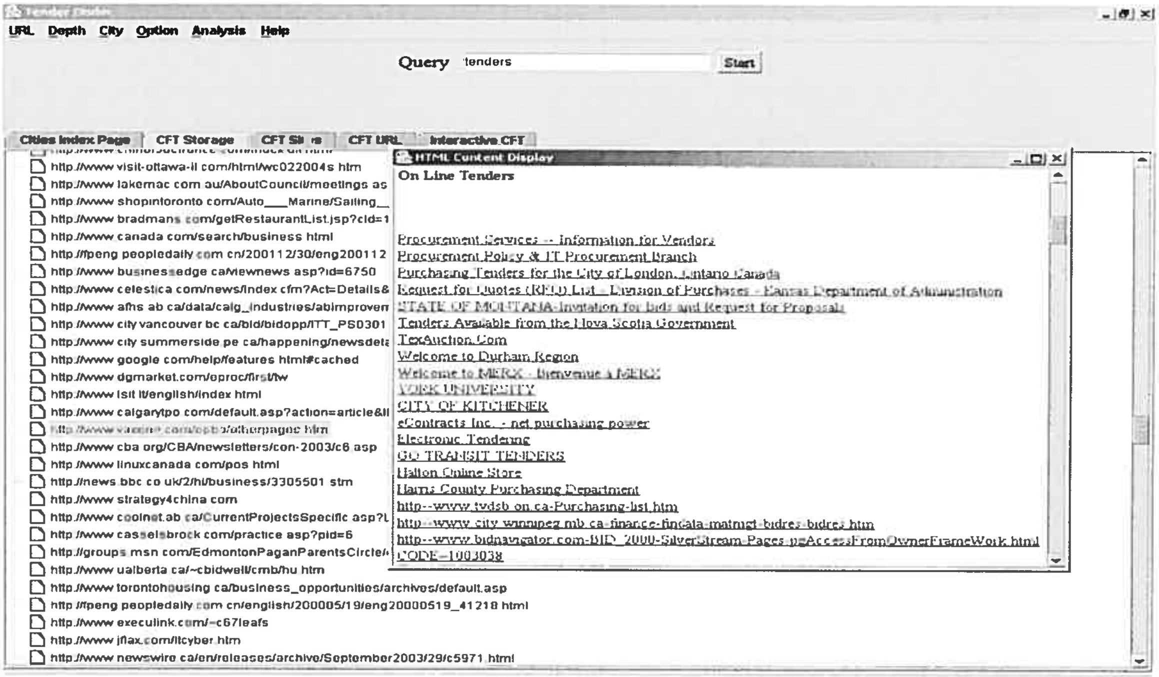
CFT Links (571)

- http://www cityofkingston ca/cityhall/tenders/eoi asp
- http://www gracekennedy com/grace/newstendernotice h
- http://www tenders net
- http://www wigwamen com/tenders/tenders html
- http://www city niagarafalls on ca/cityhall/qtenders html
- http://64 233 167 104/search?q=cache Jsiqo4pKQmsJ
- http://64 233 161 104/search?q=cache Y9kDs47d-wkJ w
- http://64 233 167 104/search?q=cache LWrlBUKnkN4J
- http://64 233 161 104/search?q=cache -Ca3YTAXQUAJ
- http://64 233 161 104/search?q=cache W6YtpbBDfW4J
- http://64 233 167 104/search?q=cache cc8WWKKsMqo
- http://www tendertouchmoving com/vehicletimes htm
- http://64 233 167 104/search?q=cache NDEUCszvYCYJ
- http://64 233 167 104/search?q=cache YdAtDhRdA2AJ v
- http://www.ottawa-airport.ca/AirportExpansion/eng/html/
- http://64 233 161 104/search?q=cache EaRSSlJuvJ0J w
- http://europa eu int/comm/economy_finance/tenders/cal
- http://64 233 167 104/search?q=cache E9gnrsHCkKAJ
- http://64 233 167 104/search?q=cache 3s4T-lzlhSUJ ww
- http://64 233 167 104/search?q=cache zHWs1hSXnsgJ
- http://www.publictender co uk
- http://www health gov au/tenders htm
- http://www.lennox-addington on ca/pagecla/Finance/Roa
- http://64 233 167 104/search?q=cache 1GBuakp3mqUJ
- http://64 233 167 104/search?q=cache pd3HwU_L4DEJ
- http://64 233 161 104/search?q=cache lzkqwrBDfawJ w
- http://www mammothweb com/lubetenders
- http://64 233 167 104/search?q=cache 6ESoM69eT8gJ

DOC #	DESCRIPTION	CLOSING DATE	STATUS	BIDDERS LIST
Quote ..2004	Installation of Overhead Doors - Fire Station #2	Closing Date/Time Thursday September 23, 2004 4:00 p m	Open	ADD ME TO THE MAILING LIST
Tender TR[illegible]2004 Title Page Drawing 1 Drawing 2 Drawing 3 Drawing 4	Landscape Development of Firemen's Park - Youth Activity Area	Closing Date/Time Friday September 17, 2004 1:15 p m Pre-Bid Meeting: A Pre-Bid Meeting is scheduled for Thursday, September 9, 2004 at 1:30 p m at the	Open	ADD ME TO THE MAILING LIST

Figure 5.3.2b *CFT Storage --- another CFT page*

The first example (Figure 5.3.2b) comes from Niagarafalls City which is a CFT.



Query tenders

- http://www visit-ottawa-il com/html/wc022004s htm
- http://www lakemac com au/AboutCouncil/meetings as
- http://www shopintoronto com/Auto___Marine/Sailing_
- http://www bradmans com/getRestaurantList.jsp?cld=1
- http://www canada com/search/business html
- http://fpeng peopledaily com cn/200112/30/eng200112
- http://www businessedge ca/viewnews asp?id=6750
- http://www celestica com/news/index cfm?Act=Details&
- http://www afhs ab ca/data/calg_industries/abimproven
- http://www city vancouver bc ca/bid/bidopp/ITT_PS0301
- http://www city summerside pe ca/happening/newsdeta
- http://www google com/help/features html#cached
- http://www dgmarket.com/eproc/first/tw
- http://www isit it/english/index html
- http://www calgarytpo com/default.asp?action=article&I
- http://www [illegible] com/[illegible]/otherpages htm
- http://www cba org/CBA/newsletters/con-2003/c6 asp
- http://www linuxcanada com/pos html
- http://news bbc co uk/2/hi/business/3305501 stm
- http://www strategy4china com
- http://www coolnet.ab ca/CurrentProjectsSpecific asp?L
- http://www casselsbrock com/practice asp?pid=6
- http://groups msn com/EdmontonPaganParentsCircle/
- http://www ualberta ca/~cbidwell/cmb/hu htm
- http://www torontohousing ca/business_opportunities/archives/default.asp
- http://fpeng peopledaily com cn/english/200005/19/eng20000519_41218 html
- http://www execulink.com/~c67leafs
- http://www jflax.com/itcyber htm
- http://www newswire ca/en/releases/archive/September2003/29/c5971 html

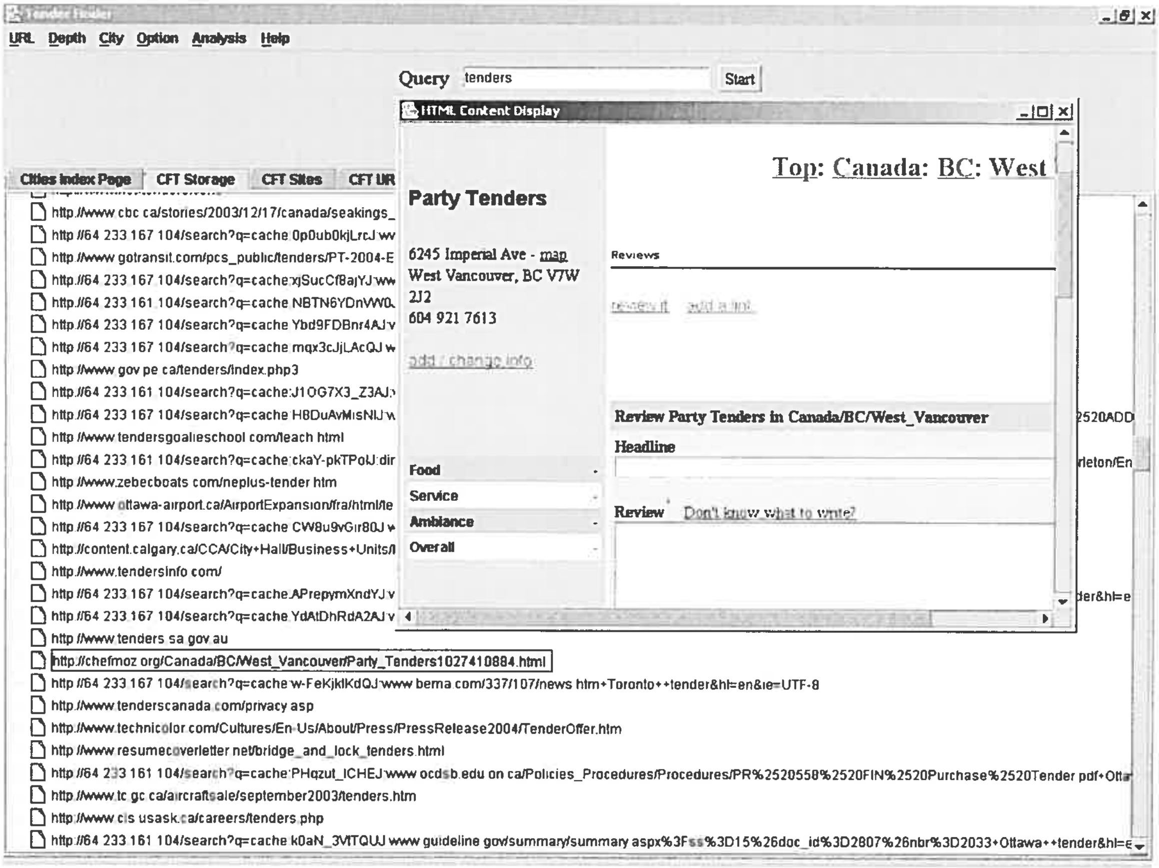
On Line Tenders

Procurement Services -- Information for Vendors
Procurement Policy & IT Procurement Branch
Purchasing Tenders for the City of London, Ontario Canada
Request for Quotes (RFQ) List - Division of Purchases - Kansas Department of Administration
STATE OF MONTANA-Invitation for bids and Request for Proposals
Tenders Available from the Nova Scotia Government
TexAuction.Com
Welcome to Durham Region
Welcome to MERX - Bienvenue à MERX
YORK UNIVERSITY
CITY OF KITCHENER
eContracts Inc. - net purchasing power
Electronic Tendering
GO TRANSIT TENDERS
Halton Online Store
Harris County Purchasing Department
http--www.tvdsb on.ca-Purchasing-list.htm
http--www.city winnipeg mb ca-finance-findata-matmgt-bidres-bidres.htm
http--www.bidnavigator.com-BID_2000-SilverStream-Pages-pgAccessFromOwnerFrameWork.html
CODE=1003038

Figure 5.3.2c *CFT Storage --- CFT pertinent page*

The second example (Figure 5.3.2c) from Vaxxine is also pertinent a CFT pertinent page (hub page).

Most of CFT storages are tender pertinent pages, this is because a part of them are collected by content lexical analysis, these pages are CFT pertinent pages with sure; others are collected by their URL strings which contain the term "tender", these pages seem to be CFT or CFT pertinent pages, but not for sure. Figure 5.3.2.d shows such an exception.



```
http.//www cbc ca/stories/2003/12/17/canada/seakings_
http //64 233 167 104/search?q=cache 0p0ub0kjLrcJ ww
http.//www gotransit.com/pcs_public/tenders/PT-2004-E
http.//64 233 167 104/search?q=cache:xjSucCf8ajYJ:ww
http.//64 233 161 104/search?q=cache NBTN6YDnVW0.
http.//64 233 167 104/search?q=cache Ybd9FDBnr4AJ:v
http //64 233 167 104/search?q=cache mqx3cJjLAcQJ w
http //www gov pe ca/tenders/index.php3
http.//64 233 161 104/search?q=cache:J1OG7X3_Z3AJ:
http.//64 233 167 104/search?q=cache H8DuAvMisNIJ:w
http //www tendersgoalieschool com/leach html
http //64 233 161 104/search?q=cache:ckaY-pkTPoIJ:dir
http //www.zebecboats com/neplus-tender htm
http.//www ottawa-airport.ca/AirportExpansion/fra/html/te
http.//64 233 167 104/search?q=cache CW8u9vGir80J w
http.//content.calgary.ca/CCA/City+Hall/Business+Units/
http.//www.tendersinfo com/
http //64 233 167 104/search?q=cache.APrepymXndYJ:v
http.//64 233 167 104/search?q=cache YdAtDhRdA2AJ:v
http //www tenders sa gov au
http.//chefmoz org/Canada/BC/West_Vancouver/Party_Tenders1027410884.html
http //64 233 167 104/search?q=cache w-FeKjklKdQJ www bema.com/337/107/news htm+Toronto++tender&hl=en&ie=UTF-8
http.//www tenderscanada com/privacy asp
http.//www.technicolor com/Cultures/En-Us/About/Press/PressRelease2004/TenderOffer.htm
http //www resumecoverletter net/bridge_and_lock_tenders.html
http //64 233 161 104/search?q=cache:PHqzut_ICHEJ www ocdsb.edu on ca/Policies_Procedures/Procedures/PR%2520558%2520FIN%2520Purchase%2520Tender pdf+Otta
http.//www.tc gc ca/aircraftsale/september2003/tenders.htm
http://www cis usask.ca/careers/tenders.php
http //64 233 161 104/search?q=cache k0aN_3VlTQUJ www guideline gov/summary/summary aspx%3Fss%3D15%26doc_id%3D2807%26nbr%3D2033+Ottawa++tender&hl=e
```

Top: Canada: BC: West

Party Tenders

6245 Imperial Ave - map
West Vancouver, BC V7W
2J2
604 921 7613

add / change info

Reviews

review it add a link

Review Party Tenders in Canada/BC/West_Vancouver

Headline

Review Don't know what to write?

Food	-
Service	-
Ambiance	-
Overall	-

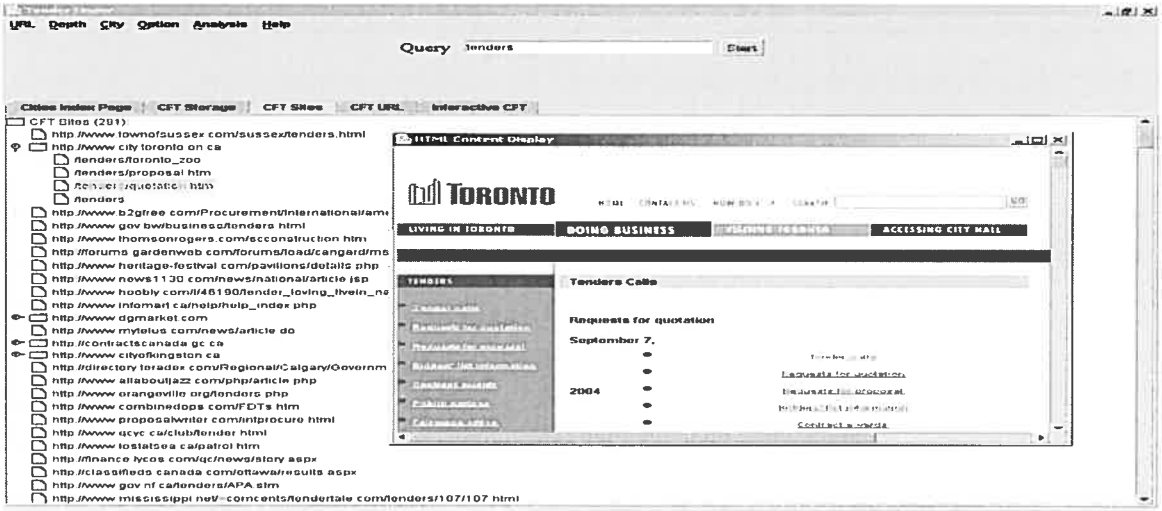
Figure 5.3.2.d CFT Storage --- not a CFT page

The last example (Figure 5.3.2d) from Vancouver is of no interest.

3). CFT sites

To do this, we reorder URLs in the CFT storage with common prefix as nodes in a tree. This function aims at finding some CFT hubs, where a lot of CFTs are gathered.

Through analyzing each candidate URL string, try to find out those CFT Hubs which are marked as middle nodes of a Jtree in figure 5.3.3.a and figure 5.3.3.b



Query: tenders

```
CFT Sites (291)
    http://www.townofsussex.com/sussex/tenders.html
    http://www.city.toronto.on.ca
        /tenders/toronto_zoo
        /tenders/proposal.htm
        /tenders/quotation.htm
        /tenders
    http://www.b2gfree.com/Procurement/International/am
    http://www.gov.bw/business/tenders.html
    http://www.thomsonrogers.com/scconstruction.htm
    http://forums.gardenweb.com/forums/load/cangard/ms
    http://www.heritage-festival.com/pavilions/details.php
    http://www.news1130.com/news/national/article.jsp
    http://www.hoobly.com/l/46190/tender_loving_livein_ne
    http://www.infomart.ca/help/help_index.php
    http://www.dgmarket.com
    http://www.mytelus.com/news/article.do
    http://contractscanada.gc.ca
    http://www.cityofkingston.ca
    http://directory.toradex.com/Regional/Calgary/Governm
    http://www.allaboutjazz.com/php/article.php
    http://www.orangeville.org/tenders.php
    http://www.combinedops.com/FDTs.htm
    http://www.proposalwriter.com/intprocure.html
    http://www.qcyc.ca/club/tender.html
    http://www.lostatsea.ca/patrol.htm
    http://finance.lycos.com/qc/news/story.aspx
    http://classifieds.canada.com/ottawa/results.aspx
    http://www.gov.nf.ca/tenders/APA.stm
    http://www.mississippi.net/~comcents/tendertale.com/tenders/107/107.html
```

Tenders Calls

Requests for quotation

September 7, 2004

- Tender calls
- Requests for quotation
- Requests for proposal
- Bidders' list information
- Contract awards

Figure 5.3.3a *CFT sites*

The site in figure 5.3.3a is http://www.city.toronto.on.ca including following leaf nodes which are all CFT pertinent. Specially, 4) is a CFT hub where we are looking for.

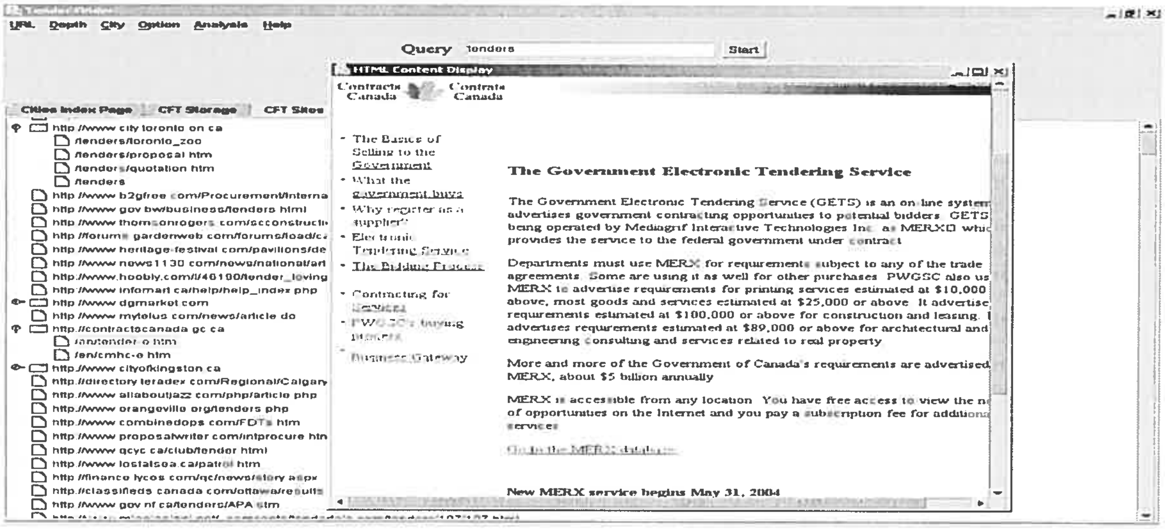
1) /tenders/toronto_zoo

2) /tenders/proposal.htm

7) /tenders/quotation.htm

8) /tenders

Again our system allows examination of the content of entries. The site in figure 5.3.3b and 5.3.3c is http://contractscanada.gc.ca, its leaf nodes are: 1)/en/tender-e.htm and 2)/en/cmhc-e.htm. 1) is an indirect CFT hub where lead to a CFT hub http://www.merx.com .



Query tenders

http //www city toronto on ca
/tenders/toronto_zoo
/tenders/proposal htm
/tenders/quotation htm
/tenders
http //www b2gfree com/Procurement/Interna
http //www gov bw/business/tenders html
http //www thomsonrogers com/scconstructi
http //forums gardenweb com/forums/load/c
http //www heritage-festival com/pavilions/de
http //www news1130 com/news/national/art
http://www.hoobly.com/l/46190/tender_loving
http //www infomart ca/help/help_index php
http //www dgmarket com
http //www mytelus com/news/article do
http //contractscanada gc ca
/en/tender-e htm
/en/cmhc-e htm
http //www cityofkingston ca
http //directory teradex com/Regional/Calgary
http //www allaboutjazz com/php/article php
http //www orangeville org/tenders php
http //www combinedops com/FDTs htm
http //www proposalwriter com/intprocure htm
http //www qcyc ca/club/tender html
http //www lostatsea ca/patrol htm
http //finance lycos com/qc/news/story aspx
http //classifieds canada com/ottawa/results
http //www gov nf ca/tenders/APA stm

- The Basics of Selling to the Government
- What the government buys
- Why register as a supplier?
- Electronic Tendering Service
- The Bidding Process
- Contracting for Services
- PWGSC's buying process
- Business Gateway

The Government Electronic Tendering Service

The Government Electronic Tendering Service (GETS) is an on line system advertises government contracting opportunities to potential bidders GETS being operated by Mediagrif Interactive Technologies Inc as MERX© whic provides the service to the federal government under contract

Departments must use MERX for requirements subject to any of the trade agreements Some are using it as well for other purchases PWGSC also us MERX to advertise requirements for printing services estimated at $10,000 above, most goods and services estimated at $25,000 or above It advertise requirements estimated at $100,000 or above for construction and leasing. advertises requirements estimated at $89,000 or above for architectural and engineering consulting and services related to real property

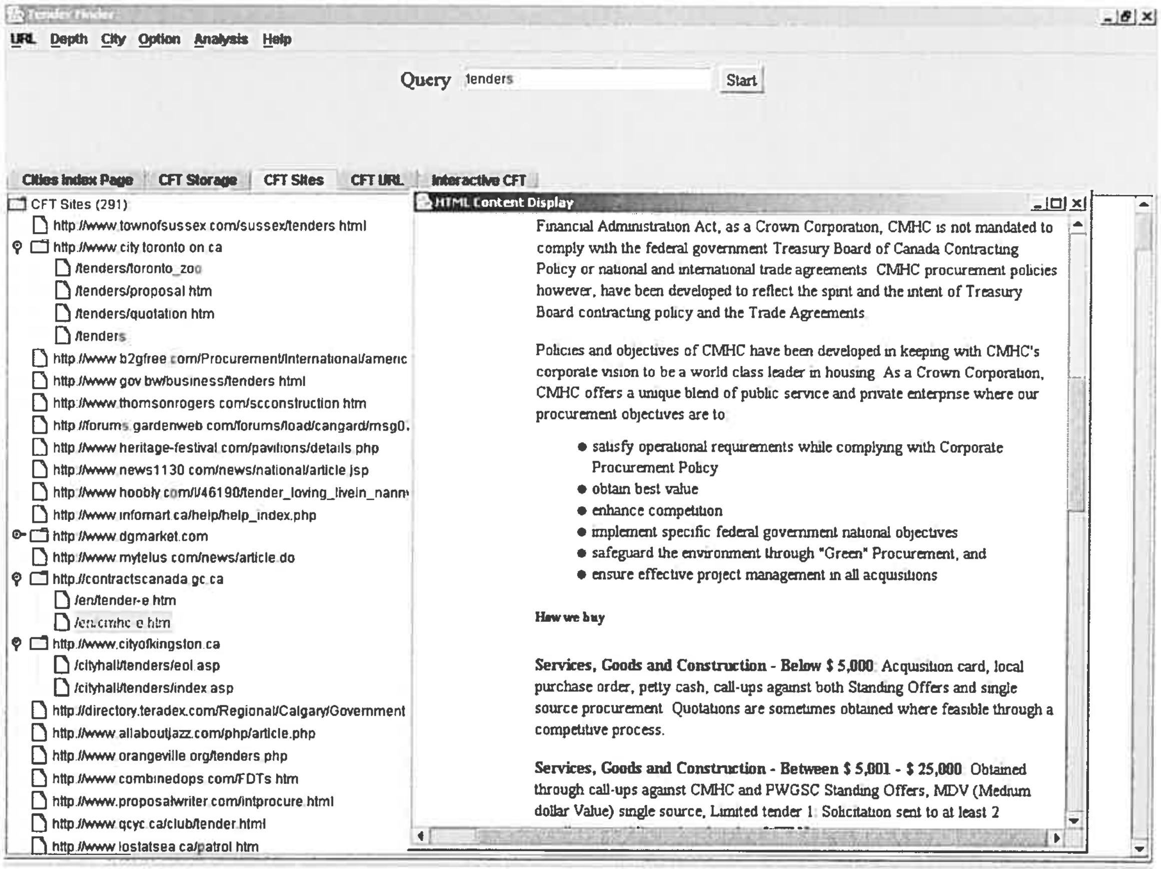
More and more of the Government of Canada's requirements are advertised MERX, about $5 billion annually

MERX is accessible from any location You have free access to view the n of opportunities on the Internet and you pay a subscription fee for additiona services

Go to the MERX database

New MERX service begins May 31, 2004

Figure 5.3.3.b *CFT sites*



Query tenders

CFT Sites (291)
http //www townofsussex com/sussex/tenders html
http //www city toronto on ca
/tenders/toronto_zoo
/tenders/proposal htm
/tenders/quotation htm
/tenders
http //www b2gfree com/Procurement/International/americ
http //www gov bw/business/tenders html
http //www thomsonrogers com/scconstruction htm
http //forums gardenweb com/forums/load/cangard/msg0.
http //www heritage-festival com/pavilions/details php
http //www news1130 com/news/national/article jsp
http //www hoobly com/l/46190/tender_loving_livein_nann
http //www infomart ca/help/help_index.php
http //www dgmarket com
http //www mytelus com/news/article do
http //contractscanada gc ca
/en/tender-e htm
/en/cmhc e htm
http //www.cityofkingston ca
/cityhall/tenders/eoi asp
/cityhall/tenders/index asp
http //directory.teradex.com/Regional/Calgary/Government
http //www allaboutjazz com/php/article.php
http //www orangeville org/tenders php
http //www combinedops com/FDTs htm
http //www.proposalwriter com/intprocure html
http //www qcyc ca/club/tender html
http //www lostatsea ca/patrol htm

Financial Administration Act, as a Crown Corporation, CMHC is not mandated to comply with the federal government Treasury Board of Canada Contracting Policy or national and international trade agreements CMHC procurement policies however, have been developed to reflect the spirit and the intent of Treasury Board contracting policy and the Trade Agreements

Policies and objectives of CMHC have been developed in keeping with CMHC's corporate vision to be a world class leader in housing As a Crown Corporation, CMHC offers a unique blend of public service and private enterprise where our procurement objectives are to

- satisfy operational requirements while complying with Corporate Procurement Policy
- obtain best value
- enhance competition
- implement specific federal government national objectives
- safeguard the environment through "Green" Procurement, and
- ensure effective project management in all acquisitions

How we buy

Services, Goods and Construction - Below $ 5,000 Acquisition card, local purchase order, petty cash, call-ups against both Standing Offers and single source procurement Quotations are sometimes obtained where feasible through a competitive process.

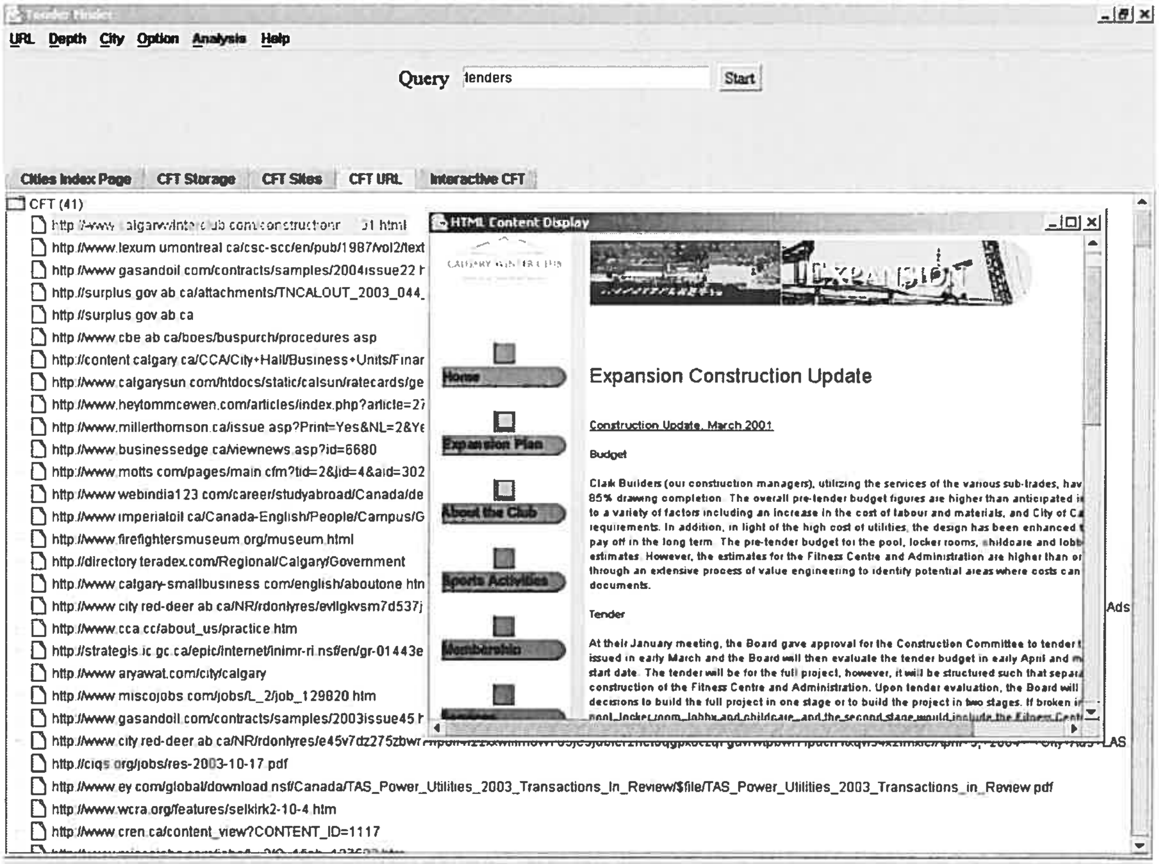
Services, Goods and Construction - Between $ 5,001 - $ 25,000 Obtained through call-ups against CMHC and PWGSC Standing Offers, MDV (Medium dollar Value) single source, Limited tender 1 Solicitation sent to at least 2

Figure 5.3.3.c *CFT sites*

Through summarizing huge amount of separate web pages, we can find some URL path where usually gather some up to date CFTs. These URL paths normally are more meaningful and stable than some separate CFTs.   Just like Figure 5.3.6.c shows.

4). CFT URL

CFT URL is a result that only based on the current searching phase. In the figure 5.3.4.a, we choose the city as Calgary first, and then query "tenders". We check the list in the figure 5.3.4.a . The figure 5.3.4a, figure 5.3.4b and figure 5.3.4c all satisfy our requirements. We'll compare the CFT URL results with Google and Yahoo in next chapter.



Query tenders

```
CFT (41)
http://www.calgarywinterclub.com/constructionmar01.html
http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/vol2/text
http://www.gasandoil.com/contracts/samples/2004issue22.h
http://surplus.gov.ab.ca/attachments/TNCALOUT_2003_044_
http://surplus.gov.ab.ca
http://www.cbe.ab.ca/boes/buspurch/procedures.asp
http://content.calgary.ca/CCA/City+Hall/Business+Units/Finar
http://www.calgarysun.com/htdocs/static/calsun/ratecards/ge
http://www.heytommcewen.com/articles/index.php?article=27
http://www.millerthomson.ca/issue.asp?Print=Yes&NL=2&Ye
http://www.businessedge.ca/viewnews.asp?id=6680
http://www.motts.com/pages/main.cfm?tid=2&jid=4&aid=302
http://www.webindia123.com/career/studyabroad/Canada/de
http://www.imperialoil.ca/Canada-English/People/Campus/G
http://www.firefightersmuseum.org/museum.html
http://directory.teradex.com/Regional/Calgary/Government
http://www.calgary-smallbusiness.com/english/aboutone.htn
http://www.city.red-deer.ab.ca/NR/rdonlyres/evilgkvsm7d537j
http://www.cca.cc/about_us/practice.htm
http://strategis.ic.gc.ca/epic/internet/inimr-ri.nsf/en/gr-01443e
http://www.aryawat.com/city/calgary
http://www.miscojobs.com/jobs/L_2/job_129820.htm
http://www.gasandoil.com/contracts/samples/2003issue45.h
http://www.city.red-deer.ab.ca/NR/rdonlyres/e45v7dz275zbwr                    LAS
http://ciqs.org/jobs/res-2003-10-17.pdf
http://www.ey.com/global/download.nsf/Canada/TAS_Power_Utilities_2003_Transactions_In_Review/$file/TAS_Power_Utilities_2003_Transactions_in_Review.pdf
http://www.wcra.org/features/selkirk2-10-4.htm
http://www.cren.ca/content_view?CONTENT_ID=1117
```

Expansion Construction Update

Construction Update, March 2001

Budget

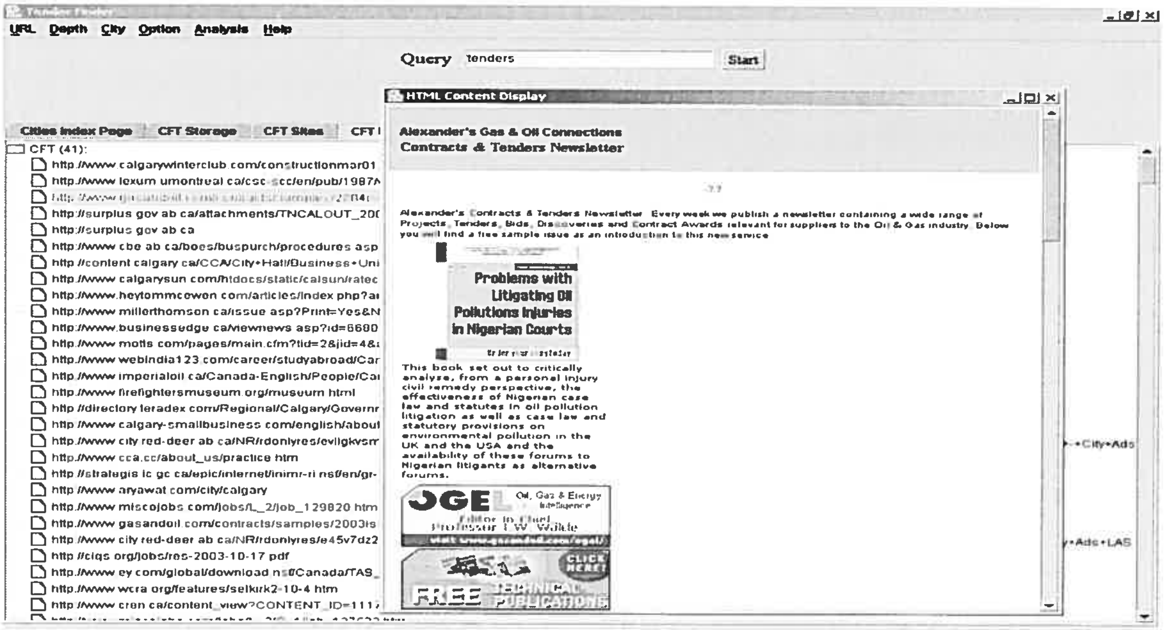
Clark Builders (our construction managers), utilizing the services of the various sub-trades, hav
85% drawing completion. The overall pre-tender budget figures are higher than anticipated i
to a variety of factors including an increase in the cost of labour and materials, and City of C
requirements. In addition, in light of the high cost of utilities, the design has been enhanced t
pay off in the long term. The pre-tender budget for the pool, locker rooms, childcare and lobb
estimates. However, the estimates for the Fitness Centre and Administration are higher than o
through an extensive process of value engineering to identify potential areas where costs can
documents.

Tender

At their January meeting, the Board gave approval for the Construction Committee to tender t
issued in early March and the Board will then evaluate the tender budget in early April and m
start date. The tender will be for the full project, however, it will be structured such that separ
construction of the Fitness Centre and Administration. Upon tender evaluation, the Board will
decisions to build the full project in one stage or to build the project in two stages. If broken i

Figure 5.3.4a   *CFT URL ---- (1)*

(1) http://www.calgarywinterclub.com/constructionmar01.html



```
CFT (41):
http://www.calgarywinterclub.com/constructionmar01
http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/
http://www.gasandoil.com/contracts/samples/2004i
http://surplus.gov.ab.ca/attachments/TNCALOUT_200
http://surplus.gov.ab.ca
http://www.cbe.ab.ca/boes/buspurch/procedures.asp
http://content.calgary.ca/CCA/City+Hall/Business+Uni
http://www.calgarysun.com/htdocs/static/calsun/ratec
http://www.heytommcewen.com/articles/index.php?ai
http://www.millerthomson.ca/issue.asp?Print=Yes&N
http://www.businessedge.ca/viewnews.asp?id=6680
http://www.motts.com/pages/main.cfm?tid=2&jid=4&
http://www.webindia123.com/career/studyabroad/Car
http://www.imperialoil.ca/Canada-English/People/Ca
http://www.firefightersmuseum.org/museum.html
http://directory.teradex.com/Regional/Calgary/Governr
http://www.calgary-smallbusiness.com/english/about
http://www.city.red-deer.ab.ca/NR/rdonlyres/evligkvsm
http://www.cca.cc/about_us/practice.htm
http://strategis.ic.gc.ca/epic/internet/inimr-ri.nsf/en/gr-
http://www.aryawat.com/city/calgary
http://www.miscojobs.com/jobs/L_2/job_129820.htm
http://www.gasandoil.com/contracts/samples/2003is
http://www.city.red-deer.ab.ca/NR/rdonlyres/e45v7dz2
http://ciqs.org/jobs/res-2003-10-17.pdf
http://www.ey.com/global/download.nsf/Canada/TAS_
http://www.wcra.org/features/selkirk2-10-4.htm
http://www.cren.ca/content_view?CONTENT_ID=111
```

Alexander's Gas & Oil Connections
Contracts & Tenders Newsletter

Alexander's Contracts & Tenders Newsletter. Every week we publish a newsletter containing a wide range of Projects, Tenders, Bids, Discoveries and Contract Awards relevant for suppliers to the Oil & Gas industry. Below you will find a free sample issue as an introduction to this new service

Problems with Litigating Oil Pollutions Injuries in Nigerian Courts

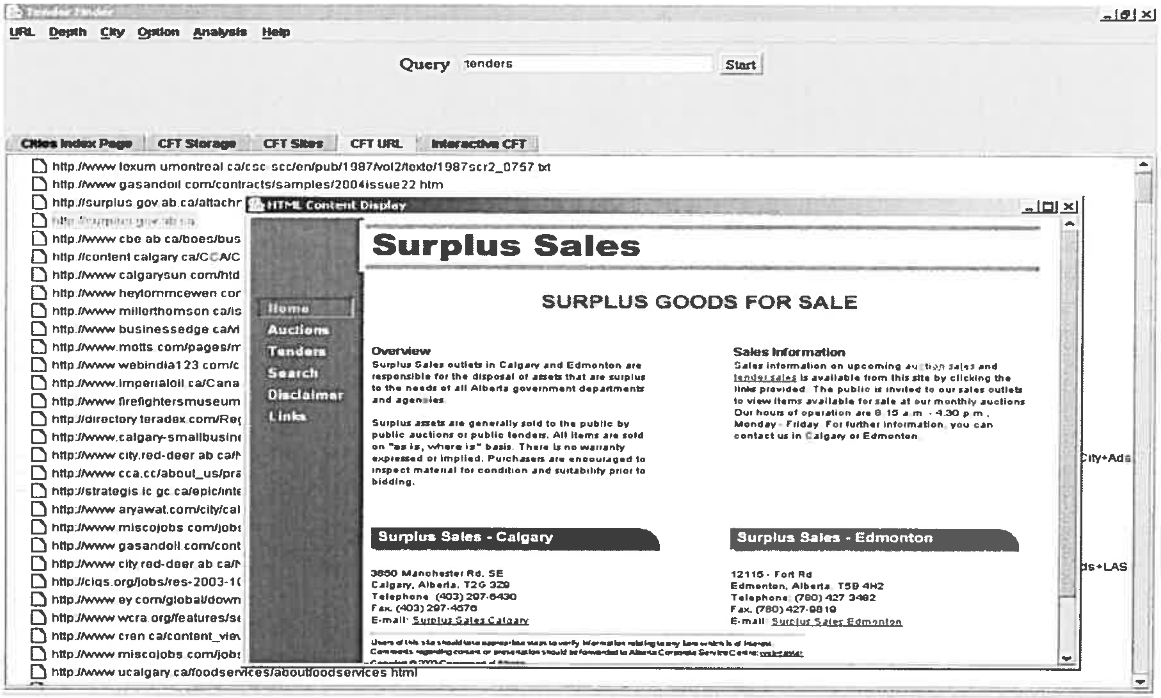
This book set out to critically analyse, from a personal injury civil remedy perspective, the effectiveness of Nigerian case law and statutes in oil pollution litigation as well as case law and statutory provisions on environmental pollution in the UK and the USA and the availability of these forums to Nigerian litigants as alternative forums.

OGEL Oil, Gas & Energy Intelligence
Editor in Chief Professor T.W. Wälde
visit www.gasandoil.com/ogel/

CLICK HERE! FREE TECHNICAL PUBLICATIONS

Figure 5.3.4b *CFT URL --- (2)*

(2) http://www.gasandoil.com/contracts/samples/2004issue22.htm



```
http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/vol2/texte/1987scr2_0757.txt
http://www.gasandoil.com/contracts/samples/2004issue22.htm
http://surplus.gov.ab.ca/attachr
http://surplus.gov.ab.ca
http://www.cbe.ab.ca/boes/bus
http://content.calgary.ca/CCA/C
http://www.calgarysun.com/htd
http://www.heytommcewen.cor
http://www.millerthomson.ca/is
http://www.businessedge.ca/vi
http://www.motts.com/pages/m
http://www.webindia123.com/c
http://www.imperialoil.ca/Cana
http://www.firefightersmuseum
http://directory.teradex.com/Re
http://www.calgary-smallbusin
http://www.city.red-deer.ab.ca/N
http://www.cca.cc/about_us/pra
http://strategis.ic.gc.ca/epic/inte
http://www.aryawat.com/city/cal
http://www.miscojobs.com/jobs
http://www.gasandoil.com/cont
http://www.city.red-deer.ab.ca/N
http://ciqs.org/jobs/res-2003-1
http://www.ey.com/global/down
http://www.wcra.org/features/se
http://www.cren.ca/content_view
http://www.miscojobs.com/jobs
http://www.ucalgary.ca/foodservices/aboutfoodservices.html
```

Surplus Sales

SURPLUS GOODS FOR SALE

Overview

Surplus Sales outlets in Calgary and Edmonton are responsible for the disposal of assets that are surplus to the needs of all Alberta government departments and agencies.

Surplus assets are generally sold to the public by public auctions or public tenders. All items are sold on "as is, where is" basis. There is no warranty expressed or implied. Purchasers are encouraged to inspect material for condition and suitability prior to bidding.

Sales Information

Sales information on upcoming auction sales and tender sales is available from this site by clicking the links provided. The public is invited to our sales outlets to view items available for sale at our monthly auctions. Our hours of operation are 8:15 a.m. - 4:30 p.m., Monday - Friday. For further information, you can contact us in Calgary or Edmonton.

Surplus Sales - Calgary

3850 Manchester Rd. SE
Calgary, Alberta, T2G 3Z9
Telephone: (403) 297-6430
Fax. (403) 297-4576
E-mail: Surplus Sales Calgary

Surplus Sales - Edmonton

12115 - Fort Rd
Edmonton, Alberta, T5B 4H2
Telephone: (780) 427 3482
Fax. (780) 427-9819
E-mail: Surplus Sales Edmonton

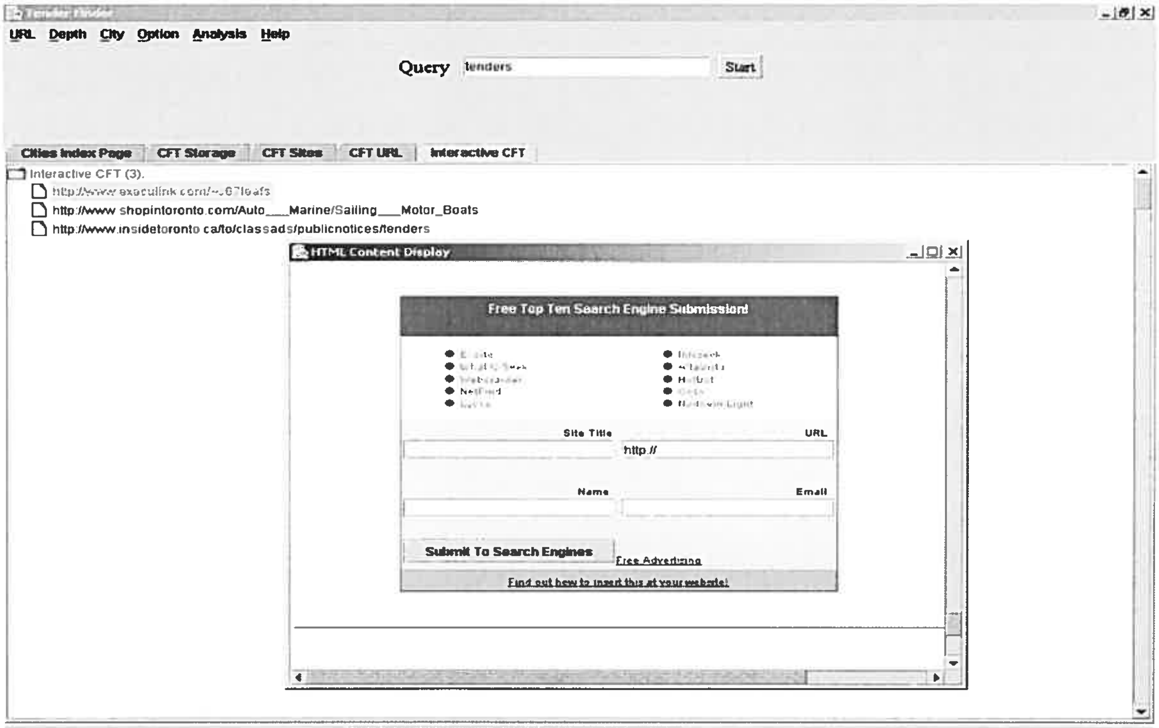
Users of this site should take appropriate steps to verify information relating to any item which is of interest.
Comments regarding content or presentation should be forwarded to Alberta Corporate Service Centre: webmaster

Figure 5.3.4c *CFT URL --- (3)*

(3) http://surplus.gov.ab.ca

5). Interactive CFT

Within the whole set of CFT pages, there are pages with various forms. The most interesting act as gateways to database of CFTs. But usually these pages have "forms" for user to communicate interactively. These are so called interactive CFT pages. To access these pages requires fill and submit forms, register as a member or create an account etc. All these steps increase the complexity for a crawler to access them automatically. Our system collects these pages, characterized as "data" pages with HTML forms. They are shown by clicking "Interactive CFT" tab. The Figure 5.3.5a shows such an interactive CFTs. collection during one search process. The Figure 5.3.5b is another example of interactive CFT formats.



Query tenders

Interactive CFT (3).

http://www.execulink.com/~067leafs

http://www.shopintoronto.com/Auto___Marine/Sailing___Motor_Boats

http://www.insidetoronto.ca/to/classads/publicnotices/tenders

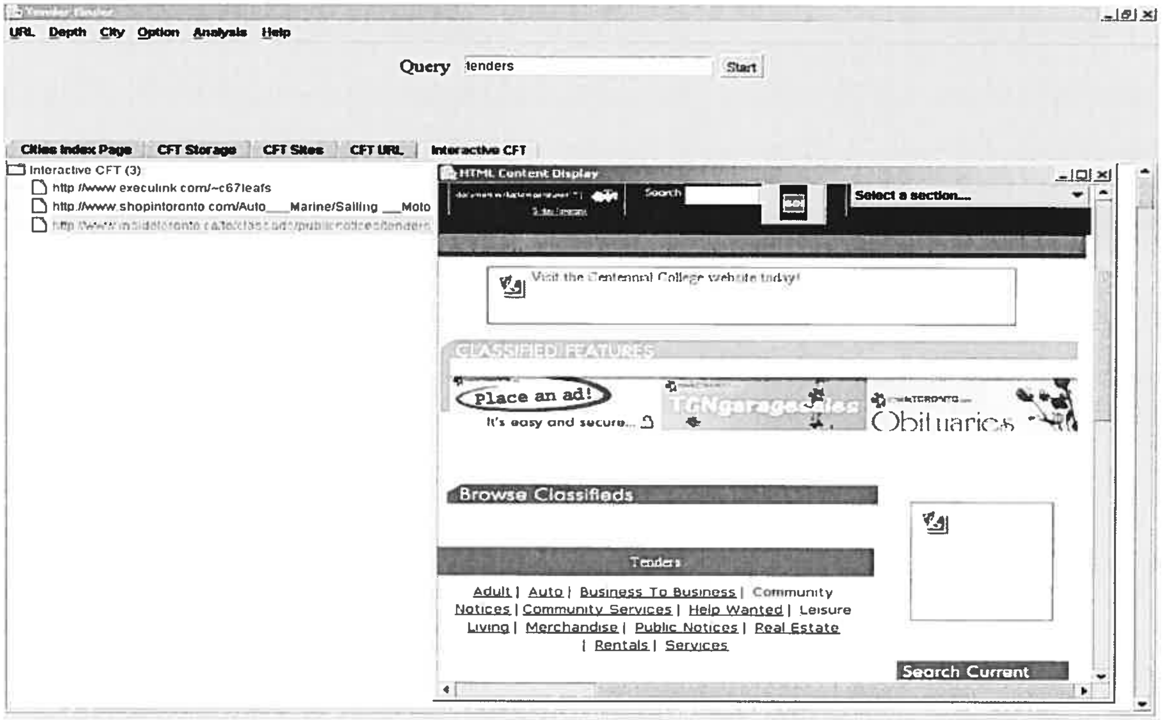
Site Title

URL http://

Name

Email

Figure 5.3.5a *Interactive CFT*



Query tenders

Interactive CFT (3)
- http //www execulink com/~c67leafs
- http //www shopintoronto com/Auto___Marine/Sailing ___Moto
- http //www insidetoronto ca/classads/publicnotices/tenders

Figure 5.3.5b *interactive CFT*

In this chapter, we introduced the system configuration, running and functions. In the following chapter we will test the system and also evaluate its general performances by comparing with the CSE the system based on.

Chapter 6
Testing and Evaluation

In this chapter we'll test the domain specific system with some sets of web pages. Testing will cover system tools test such as the structural test, the lexical testing as well as system function test including CFT storage, CFT sites and CFT URL etc.

6.1 System tools test

To evaluate web pages, our system also provides two system tools to support the analysis of web pages. These two tools are structural analysis and lexical analysis.

6.1.1 Structural analysis

The structural analysis focuses on extracting a text area from a web page with a complex logical structure. Considering most complex web pages are composed with tables within tables, this extracting process sets out from decomposing a web page to several tables and then analyzes these tables. We choose two examples web page for structural analysis testing. The examples are:

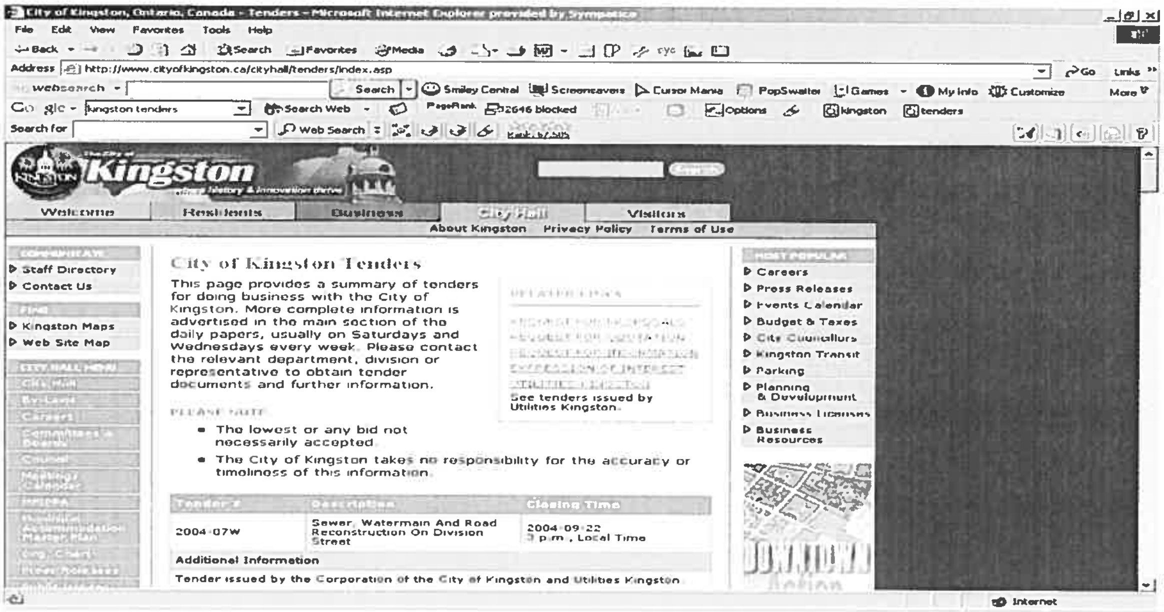
1) http://www.cityofkingston.ca/cityhall/tenders/index.asp
2) http://www.tc.gc.ca/aircraftsale/tender.htm

Test 1:

The page concerns Kingston * looks like figure 6.1.1a and figure 6.1.1b in next page. Previously, in section 5.4, we used this page to show our structural analysis tool.

* http://www.cityofkingston.ca/cityhall/tenders/index.asp

Figure 5.3.6 showed that it was composed of about 30 HTML tables and table 23 was judged to contain the most pertinent data. Figure 6.1.1c and Figure 6.1.1d show this part of the page in more detail.



City of Kingston Tenders

This page provides a summary of tenders for doing business with the City of Kingston. More complete information is advertised in the main section of the daily papers, usually on Saturdays and Wednesdays every week. Please contact the relevant department, division or representative to obtain tender documents and further information.

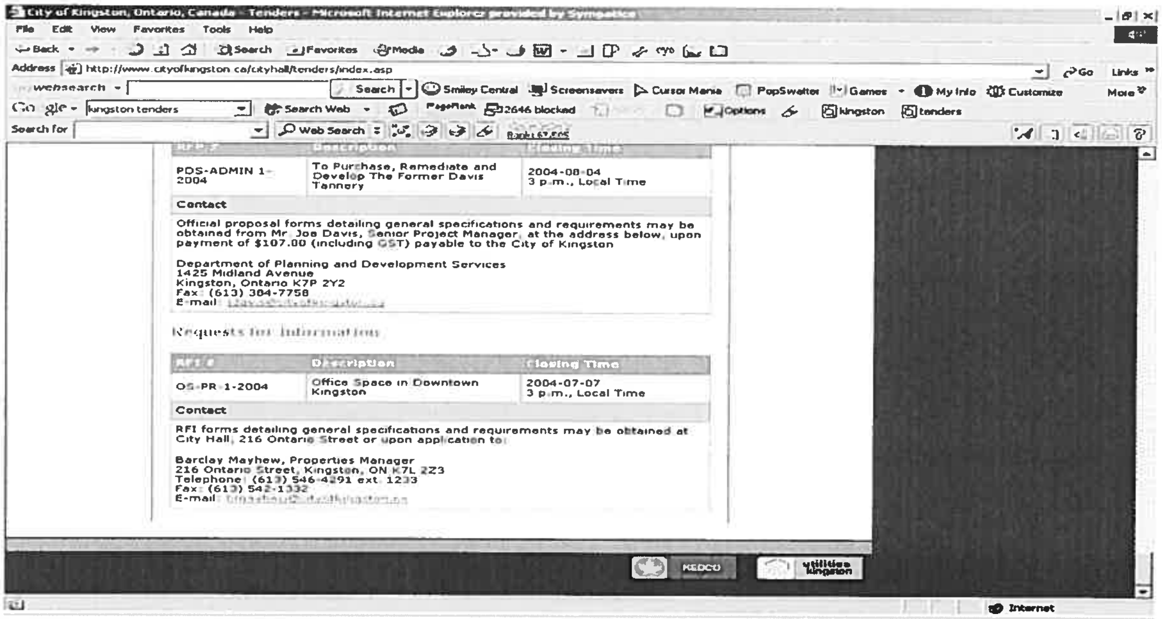
See tenders issued by Utilities Kingston.

PLEASE NOTE

- The lowest or any bid not necessarily accepted.
- The City of Kingston takes no responsibility for the accuracy or timeliness of this information.

Tender #	Description	Closing Time
2004-07W	Sewer, Watermain And Road Reconstruction On Division Street	2004-09-22 3 p.m., Local Time
Additional Information		
Tender issued by the Corporation of the City of Kingston and Utilities Kingston		

Figure 6.1.1a *Mixed page top side*


City of Kingston, Ontario, Canada - Tenders - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Address http://www.cityofkingston.ca/cityhall/tenders/index.asp
kingston tenders
KEDCO
utilities kingston
Internet

RFP #	Description	Closing Time
PDS-ADMIN 1-2004	To Purchase, Remediate and Develop The Former Davis Tannery	2004-08-04 3 p.m., Local Time
Contact		

Official proposal forms detailing general specifications and requirements may be obtained from Mr. Joe Davis, Senior Project Manager, at the address below, upon payment of $107.00 (including GST) payable to the City of Kingston

Department of Planning and Development Services
1425 Midland Avenue
Kingston, Ontario K7P 2Y2
Fax: (613) 384-7758
E-mail: [illegible]

Requests for Information

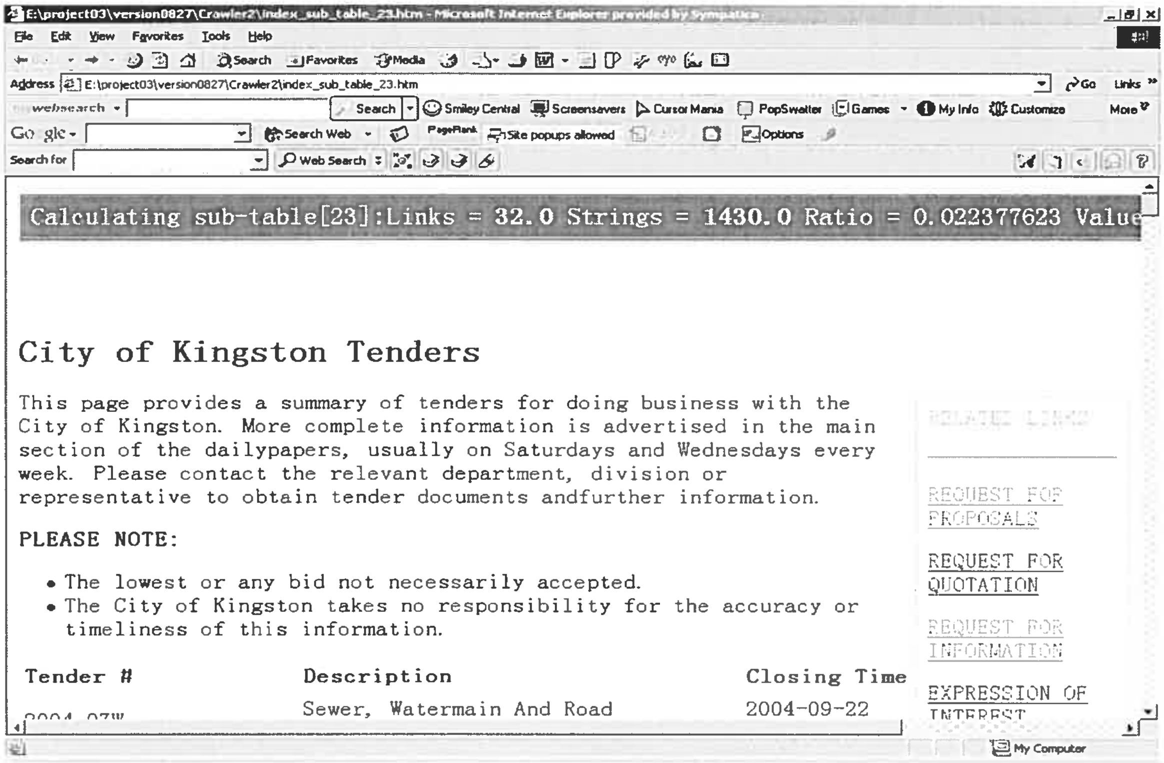
RFI #	Description	Closing Time
OS-PR-1-2004	Office Space in Downtown Kingston	2004-07-07 3 p.m., Local Time
Contact		

RFI forms detailing general specifications and requirements may be obtained at City Hall, 216 Ontario Street or upon application to:

Barclay Mayhew, Properties Manager
216 Ontario Street, Kingston, ON K7L 2Z3
Telephone: (613) 546-4291 ext. 1233
Fax: (613) 542-1332
E-mail: [illegible]

Figure 6.1.1b *Mixed page bottom side*

After extracting:


E:\project03\version0827\Crawler2\index_sub_table_23.htm - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Search Favorites Media
Address E:\project03\version0827\Crawler2\index_sub_table_23.htm
Go Links
websearch Search Smiley Central Screensavers Cursor Mania PopSwatter Games My Info Customize More
Google Search Web PageRank Site popups allowed Options
Search for Web Search
My Computer

Calculating sub-table[23]:Links = 32.0 Strings = 1430.0 Ratio = 0.022377623 Value

City of Kingston Tenders

This page provides a summary of tenders for doing business with the City of Kingston. More complete information is advertised in the main section of the dailypapers, usually on Saturdays and Wednesdays every week. Please contact the relevant department, division or representative to obtain tender documents andfurther information.

PLEASE NOTE:

- The lowest or any bid not necessarily accepted.
- The City of Kingston takes no responsibility for the accuracy or timeliness of this information.

Tender #	Description	Closing Time
2004-07W	Sewer, Watermain And Road	2004-09-22

RELATED LINKS

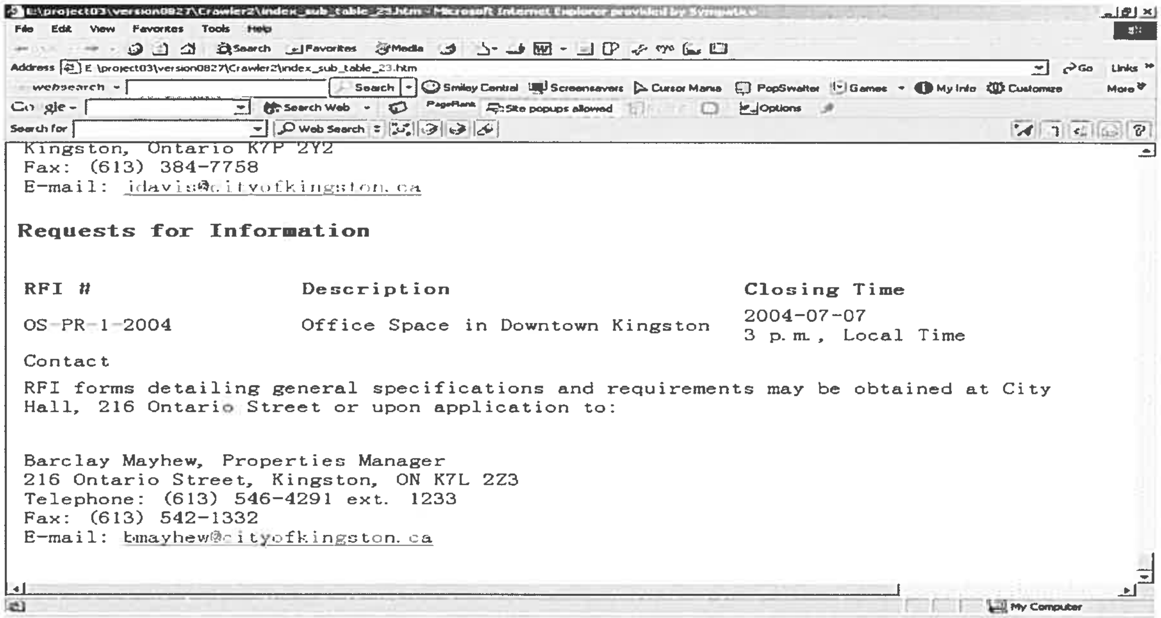
REQUEST FOR PROPOSALS

REQUEST FOR QUOTATION

REQUEST FOR INFORMATION

EXPRESSION OF INTEREST

Figure 6.1.1c *Extracted sub-table top side*


E:\project03\version0827\Crawler2\index_sub_table_23.htm - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Search Favorites Media
Address E:\project03\version0827\Crawler2\index_sub_table_23.htm
Go Links
websearch Search Smiley Central Screensavers Cursor Mania PopSwatter Games My Info Customize More
Google Search Web PageRank Site popups allowed Options
Search for Web Search
My Computer

Kingston, Ontario K7P 2Y2
Fax: (613) 384-7758
E-mail: idavis@cityofkingston.ca

Requests for Information

RFI #	Description	Closing Time
OS-PR-1-2004	Office Space in Downtown Kingston	2004-07-07 3 p.m., Local Time

Contact

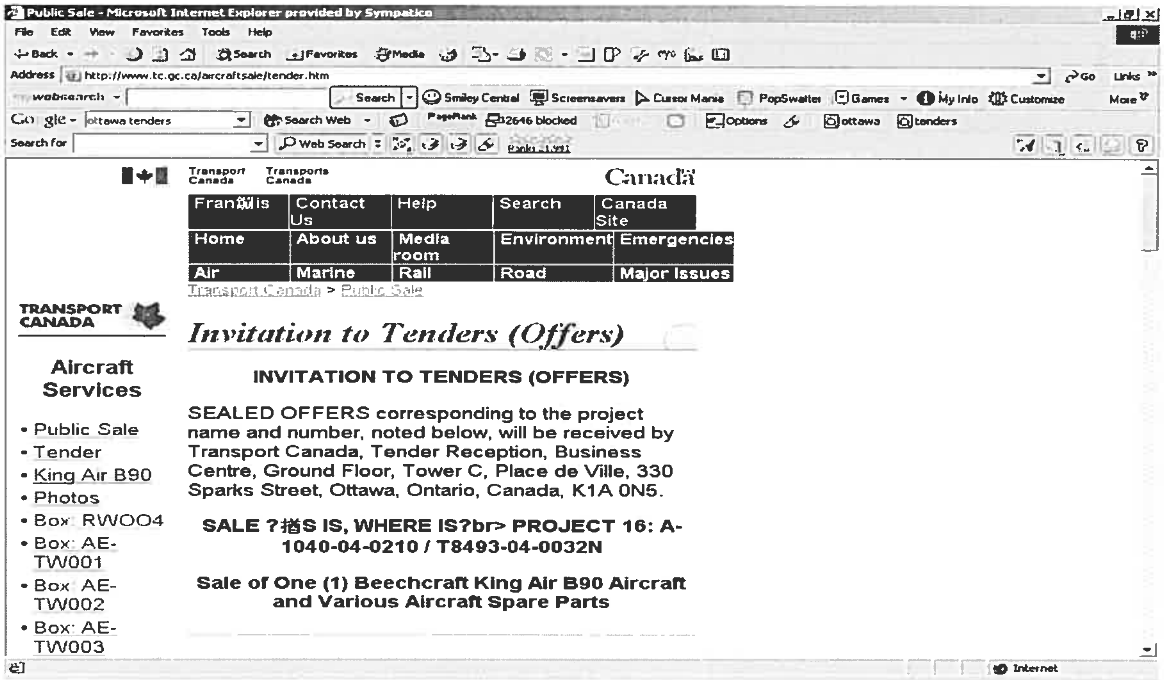
RFI forms detailing general specifications and requirements may be obtained at City Hall, 216 Ontario Street or upon application to:

Barclay Mayhew, Properties Manager
216 Ontario Street, Kingston, ON K7L 2Z3
Telephone: (613) 546-4291 ext. 1233
Fax: (613) 542-1332
E-mail: bmayhew@cityofkingston.ca

Figure 6.1.1d *Extracted sub-table bottom side*

Test 2:

The original web page looks like figure6.1.2a and extracted one like figure 6.1.2b shown. There are not much differences after extracting.


File Edit View Favorites Tools Help
Back Search Favorites Media
Go Links
websearch Search Smiley Central Screensavers Cursor Mania PopSwatter Games My Info Customize More
Google Search Web PageRank 2646 blocked Options ottawa tenders
Search for Web Search
Internet

Public Sale - Microsoft Internet Explorer provided by Sympatico

Address http://www.tc.gc.ca/aircraftsale/tender.htm

Google: ottawa tenders

Transport Canada Transports Canada Canadä

Fran嘛is	Contact Us	Help	Search	Canada Site
Home	About us	Media room	Environment	Emergencies
Air	Marine	Rail	Road	Major Issues

Transport Canada > Public Sale

TRANSPORT CANADA

Invitation to Tenders (Offers)

Aircraft Services

- Public Sale
- Tender
- King Air B90
- Photos
- Box: RWOO4
- Box: AE-TW001
- Box: AE-TW002
- Box: AE-TW003

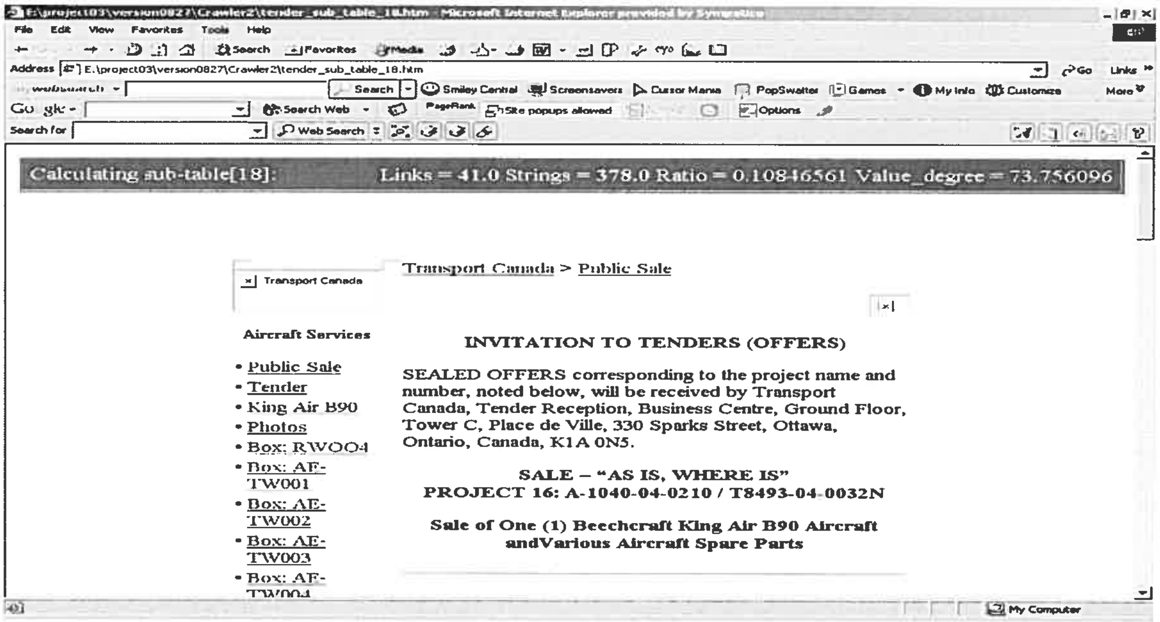
INVITATION TO TENDERS (OFFERS)

SEALED OFFERS corresponding to the project name and number, noted below, will be received by Transport Canada, Tender Reception, Business Centre, Ground Floor, Tower C, Place de Ville, 330 Sparks Street, Ottawa, Ontario, Canada, K1A 0N5.

SALE ?揩S IS, WHERE IS?br> PROJECT 16: A-1040-04-0210 / T8493-04-0032N

Sale of One (1) Beechcraft King Air B90 Aircraft and Various Aircraft Spare Parts

Figure 6.1.2a *Original web page top side*


File Edit View Favorites Tools Help
Search Favorites Media
Go Links
websearch Search Smiley Central Screensavers Cursor Mania PopSwatter Games My Info Customize More
Google Search Web PageRank Site popups allowed Options
Search for Web Search
My Computer

E:\project03\version0827\Crawler2\tender_sub_table_18.htm - Microsoft Internet Explorer provided by Sympatico

Address E:\project03\version0827\Crawler2\tender_sub_table_18.htm

Calculating sub-table[18]: Links = 41.0 Strings = 378.0 Ratio = 0.10846561 Value_degree = 73.756096

Transport Canada

Transport Canada > Public Sale

Aircraft Services

- Public Sale
- Tender
- King Air B90
- Photos
- Box: RWOO4
- Box: AE-TW001
- Box: AE-TW002
- Box: AE-TW003
- Box: AE-TW004

INVITATION TO TENDERS (OFFERS)

SEALED OFFERS corresponding to the project name and number, noted below, will be received by Transport Canada, Tender Reception, Business Centre, Ground Floor, Tower C, Place de Ville, 330 Sparks Street, Ottawa, Ontario, Canada, K1A 0N5.

SALE – "AS IS, WHERE IS"
PROJECT 16: A-1040-04-0210 / T8493-04-0032N

Sale of One (1) Beechcraft King Air B90 Aircraft andVarious Aircraft Spare Parts

Figure 6.1.2b *Extracted sub-table top side*

These results show that the structural analysis is especially proper for those pages which contain a complex logical structure.

6.1.2 Lexical analysis

The lexical analysis provides for both web files and local directory/files. The analysis can also accumulate the words and frequencies during an analyzing session. Therefore, we can use it to analyze a group of web pages.

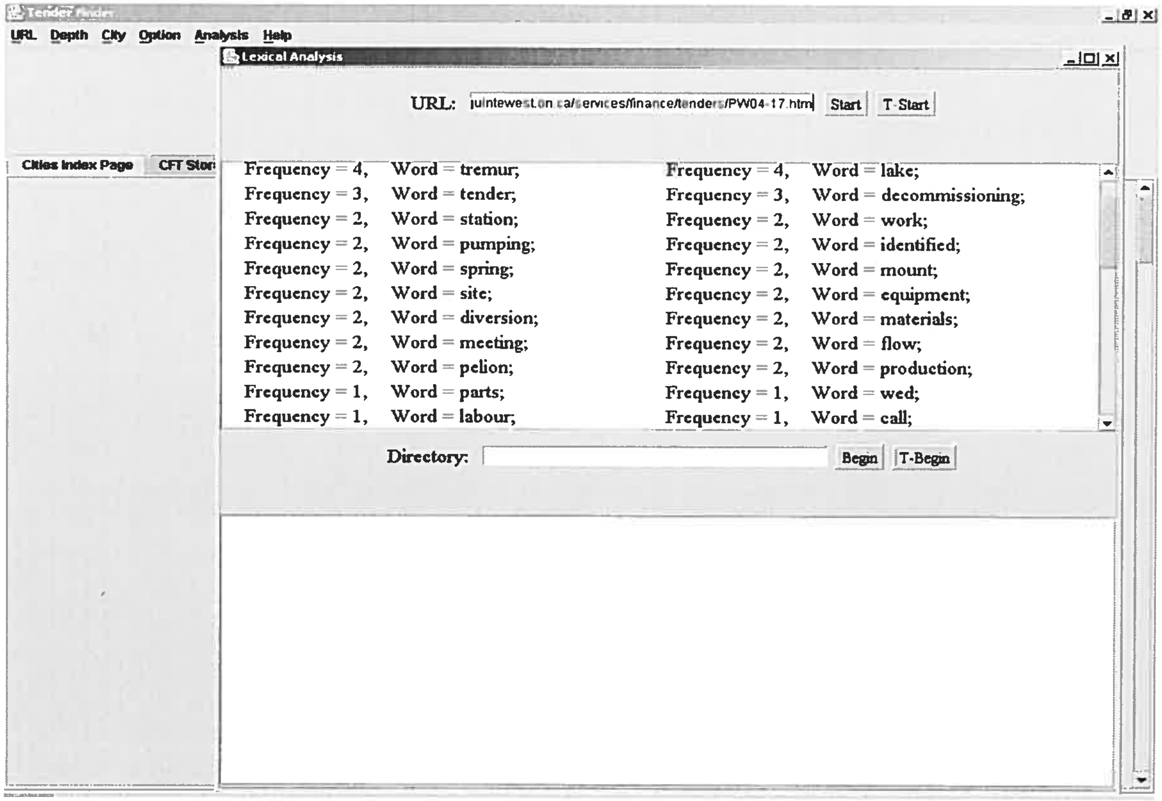
For example, we have 2 web pages:

1)Edmonton City:

http://www.city.quintewest.on.ca/services/finance/tenders/PW04-17.htm

2) St John City: http://www.stjohns.ca/csj/TendersDetails?id=156&loc=W

add them one by one to lexical analysis, we get the result as following:

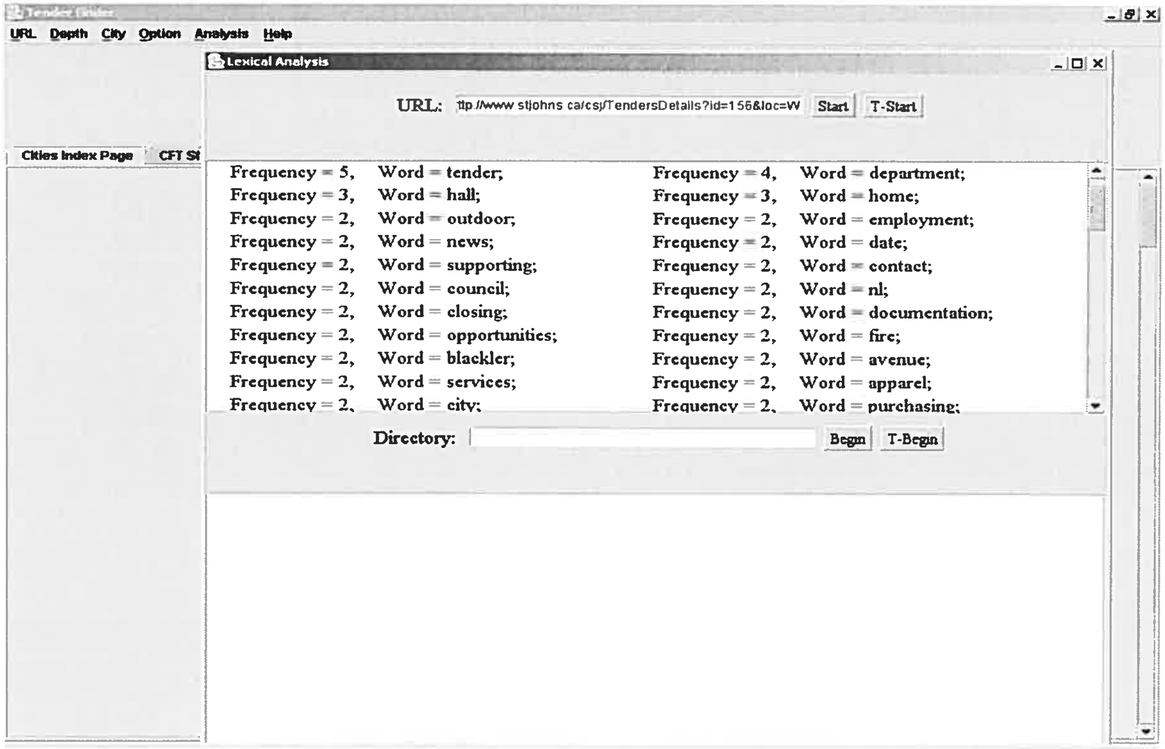


URL: juintewest.on.ca/services/finance/tenders/PW04-17.htm

```
Frequency = 4,   Word = tremur;        Frequency = 4,   Word = lake;
Frequency = 3,   Word = tender;        Frequency = 3,   Word = decommissioning;
Frequency = 2,   Word = station;       Frequency = 2,   Word = work;
Frequency = 2,   Word = pumping;       Frequency = 2,   Word = identified;
Frequency = 2,   Word = spring;        Frequency = 2,   Word = mount;
Frequency = 2,   Word = site;          Frequency = 2,   Word = equipment;
Frequency = 2,   Word = diversion;     Frequency = 2,   Word = materials;
Frequency = 2,   Word = meeting;       Frequency = 2,   Word = flow;
Frequency = 2,   Word = pelion;        Frequency = 2,   Word = production;
Frequency = 1,   Word = parts;         Frequency = 1,   Word = wed;
Frequency = 1,   Word = labour;        Frequency = 1,   Word = call;
```

Directory:

Figure 6.1.3a *Page 1) lexical analysis*

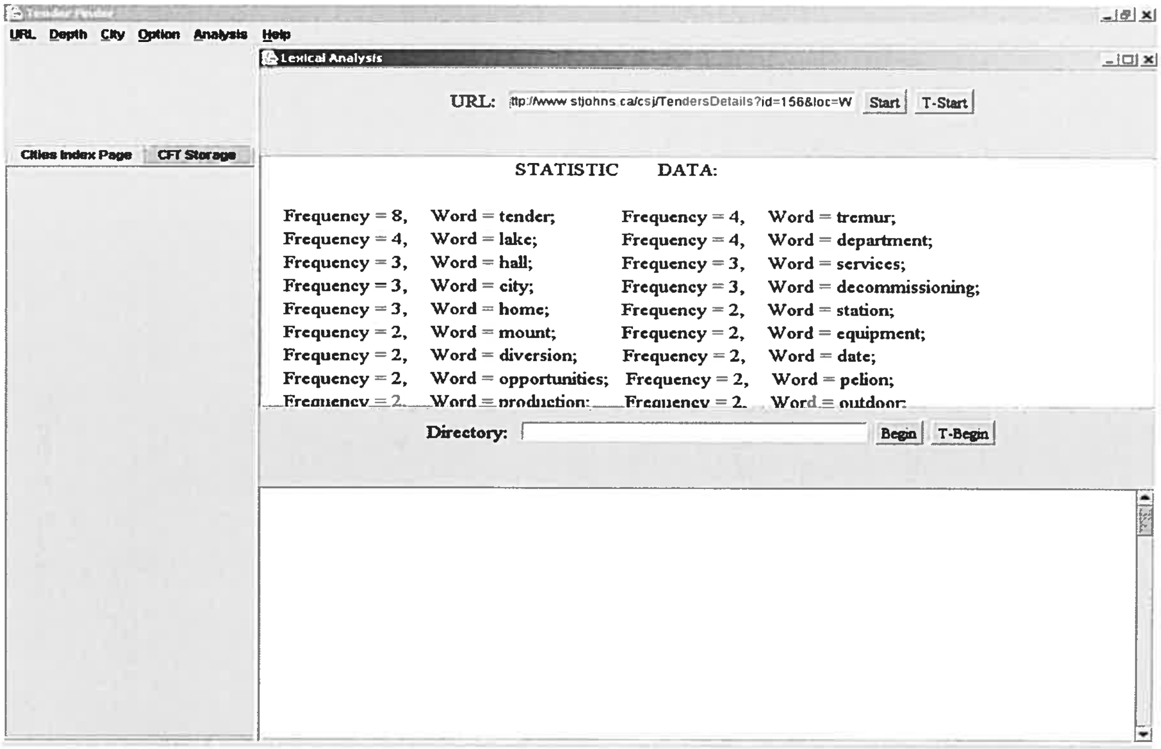


URL: ttp://www stjohns ca/csj/TendersDetails?id=156&loc=W

```
Frequency = 5,   Word = tender;            Frequency = 4,   Word = department;
Frequency = 3,   Word = hall;              Frequency = 3,   Word = home;
Frequency = 2,   Word = outdoor;           Frequency = 2,   Word = employment;
Frequency = 2,   Word = news;              Frequency = 2,   Word = date;
Frequency = 2,   Word = supporting;        Frequency = 2,   Word = contact;
Frequency = 2,   Word = council;           Frequency = 2,   Word = nl;
Frequency = 2,   Word = closing;           Frequency = 2,   Word = documentation;
Frequency = 2,   Word = opportunities;     Frequency = 2,   Word = fire;
Frequency = 2,   Word = blackler;          Frequency = 2,   Word = avenue;
Frequency = 2,   Word = services;          Frequency = 2,   Word = apparel;
Frequency = 2,   Word = city;              Frequency = 2,   Word = purchasing;
```

Directory:

Figure 6.1.3b *Page 2) lexical analysis*


Tender Finder
URL Depth City Option Analysis Help
Lexical Analysis
Start T-Start
Cities Index Page CFT Storage
Begin T-Begin

URL: ttp://www stjohns ca/csj/TendersDetails?id=156&loc=W

```
                    STATISTIC      DATA:

Frequency = 8,   Word = tender;          Frequency = 4,   Word = tremur;
Frequency = 4,   Word = lake;            Frequency = 4,   Word = department;
Frequency = 3,   Word = hall;            Frequency = 3,   Word = services;
Frequency = 3,   Word = city;            Frequency = 3,   Word = decommissioning;
Frequency = 3,   Word = home;            Frequency = 2,   Word = station;
Frequency = 2,   Word = mount;           Frequency = 2,   Word = equipment;
Frequency = 2,   Word = diversion;       Frequency = 2,   Word = date;
Frequency = 2,   Word = opportunities;   Frequency = 2,   Word = pelion;
Frequency = 2,   Word = production;      Frequency = 2,   Word = outdoor;
```

Directory:

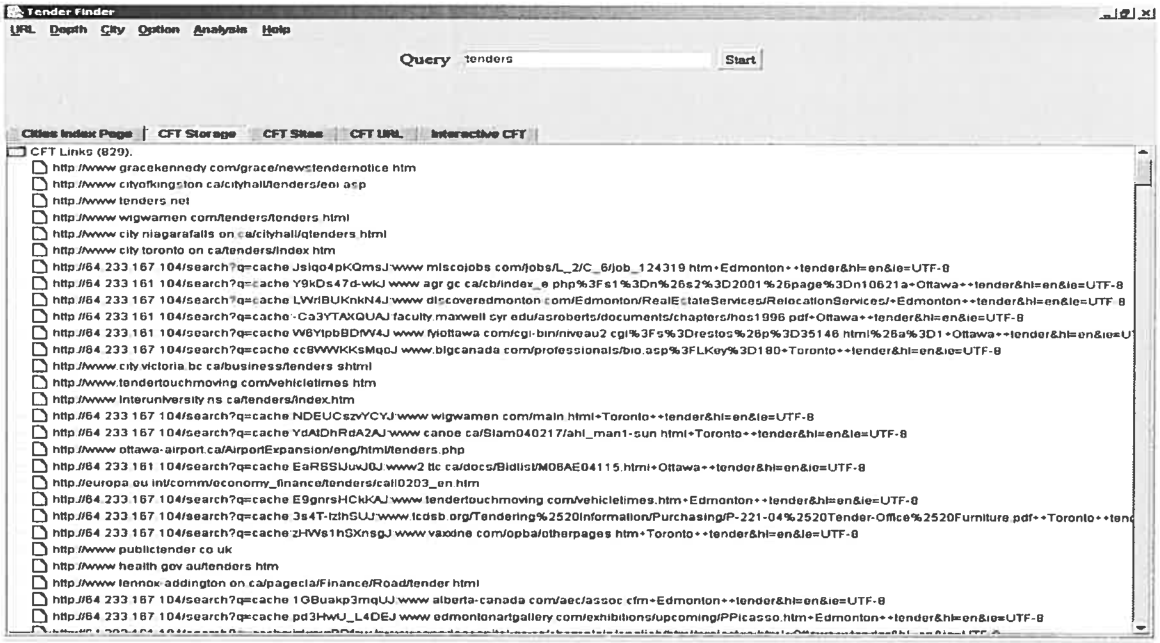
Figure 6.1.3c. *Page 1) & 2) lexical analysis*

From the results we can see that the 5 most high frequency words in first page are: tremur (4), lake (4), tender (3), decommissioning (3) and station (2). For the second page these high frequency words are tender (5), department (4), hall (3), home (3), and lot of words with frequency equal 2, such as outdoor, employment, news, date, city etc. The figure 6.1.3.c shows the cumulative frequencies for the two pages. The first 5 high frequency words are: tender (8), trmur(4), lake(4), department(4), hall(3), services(3), city(3), decommissioning (3), home(3), station(2) etc. So, if we want to test a set of web page, just add them to the URL field and press start button.

6.2 System test

To test the system functions, we first configure the system parameters. For example, we configure page_number=1, crawl_depth=2; city=Calgary; no tender extension and no tender filter.

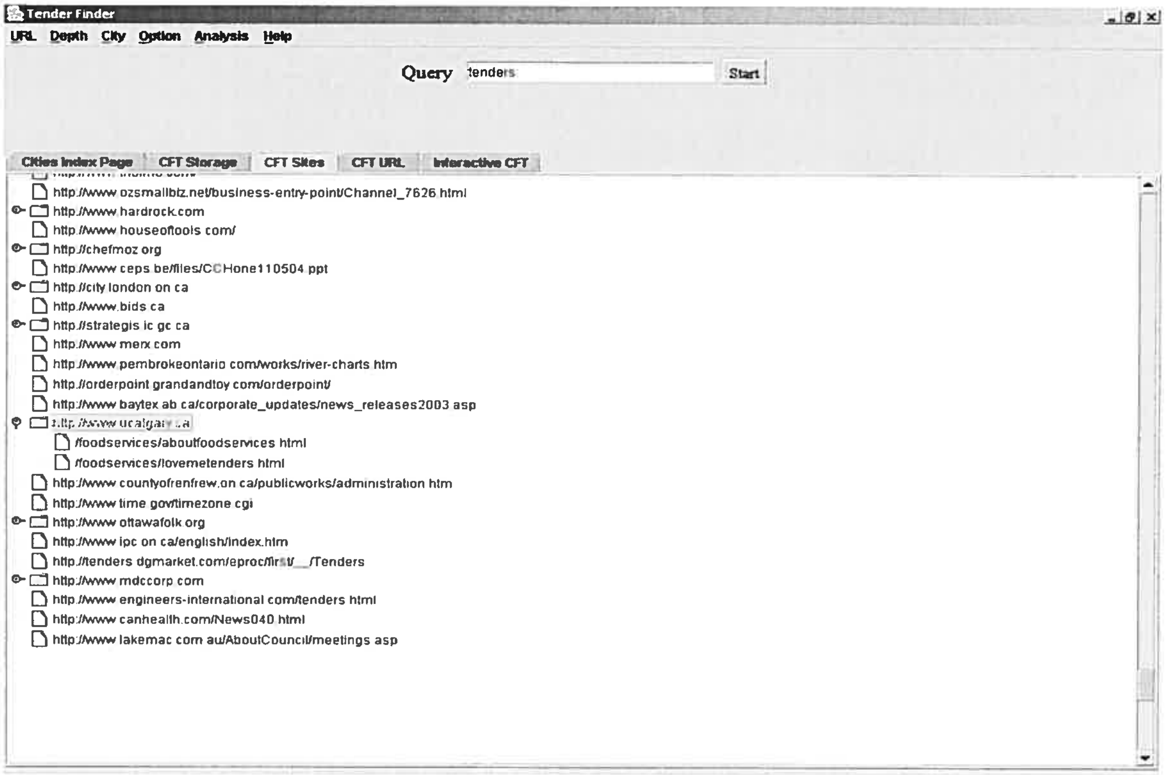
6.2.1 CFT storage and CFT sites



Query tenders

```
CFT Links (829).
  http://www gracekennedy com/grace/newstendernotice htm
  http://www cityofkingston ca/cityhall/tenders/eoi asp
  http://www tenders net
  http://www wigwamen com/tenders/tenders html
  http://www city niagarafalls on ca/cityhall/qtenders html
  http://www city toronto on ca/tenders/index htm
  http://64 233 167 104/search?q=cache Jsiqo4pKQmsJ www miscojobs com/jobs/L_2/C_6/job_124319 htm+Edmonton++tender&hl=en&ie=UTF-8
  http://64 233 161 104/search?q=cache Y9kDs47d-wkJ www agr gc ca/cb/index_e php%3Fs1%3Dn%26s2%3D2001%26page%3Dn10621a+Ottawa++tender&hl=en&ie=UTF-8
  http://64 233 167 104/search?q=cache LWrIBUKnkN4J www discoveredmonton com/Edmonton/RealEstateServices/RelocationServices/+Edmonton++tender&hl=en&ie=UTF 8
  http://64 233 161 104/search?q=cache -Ca3YTAXQUAJ faculty maxwell syr edu/asroberts/documents/chapters/hos1996 pdf+Ottawa++tender&hl=en&ie=UTF-8
  http://64 233 161 104/search?q=cache W6YlpbBDfW4J www fyiottawa com/cgi-bin/niveau2 cgi%3Fs%3Drestos%26p%3D35146 html%26a%3D1+Ottawa++tender&hl=en&ie=U
  http://64 233 167 104/search?q=cache cc8WWKKsMqoJ www.bigcanada com/professionals/bio.asp%3FLKey%3D180+Toronto++tender&hl=en&ie=UTF-8
  http://www.city victoria bc ca/business/tenders shtml
  http://www.tendertouchmoving com/vehicletimes htm
  http://www interuniversity ns ca/tenders/index.htm
  http://64 233 167 104/search?q=cache NDEUCszvYCYJ www wigwamen com/main html+Toronto++tender&hl=en&ie=UTF-8
  http://64 233 167 104/search?q=cache YdAtDhRdA2AJ www canoe ca/Slam040217/ahl_man1-sun html+Toronto++tender&hl=en&ie=UTF-8
  http://www ottawa-airport.ca/AirportExpansion/eng/html/tenders.php
  http://64 233 161 104/search?q=cache EaRSSIJuwJ0J www2 tc ca/docs/Bidlist/M06AE04115.html+Ottawa++tender&hl=en&ie=UTF-8
  http://europa eu int/comm/economy_finance/tenders/call0203_en.htm
  http://64 233 167 104/search?q=cache E9gnrsHCkKAJ www tendertouchmoving com/vehicletimes.htm+Edmonton++tender&hl=en&ie=UTF-8
  http://64 233 167 104/search?q=cache 3s4T-tzlhSUJ www.tcdsb.org/Tendering%2520Information/Purchasing/P-221-04%2520Tender-Office%2520Furniture pdf++Toronto++tend
  http://64 233 167 104/search?q=cache zHWs1hSXnsgJ www vaxxine com/opba/otherpages htm+Toronto++tender&hl=en&ie=UTF-8
  http://www publictender co uk
  http://www health gov au/tenders htm
  http://www lennox-addington on ca/pagecla/Finance/Road/tender html
  http://64 233 167 104/search?q=cache 1GBuakp3mqUJ www alberta-canada com/aec/assoc cfm+Edmonton++tender&hl=en&ie=UTF-8
  http://64 233 167 104/search?q=cache pd3HwU_L4DEJ www edmontonartgallery com/exhibitions/upcoming/PPicasso.htm+Edmonton++tender&hl=en&ie=UTF-8
```

Figure 6.1.4a *CFT storage -test*



Query: tenders

```
  http://www ozsmallbiz.net/business-entry-point/Channel_7626 html
  http://www hardrock.com
  http://www houseoftools com/
  http://chefmoz org
  http://www ceps be/files/CCHone110504 ppt
  http://city london on ca
  http://www.bids ca
  http://strategis ic gc ca
  http://www merx com
  http://www pembrokeontario com/works/river-charts htm
  http://orderpoint grandandtoy com/orderpoint/
  http://www baytex ab ca/corporate_updates/news_releases2003 asp
  http://www ucalgary ca
      /foodservices/aboutfoodservices html
      /foodservices/lovemetenders html
  http://www countyofrenfrew.on ca/publicworks/administration htm
  http://www time gov/timezone cgi
  http://www ottawafolk org
  http://www ipc on ca/english/index.htm
  http://tenders dgmarket.com/eproc/first/__/Tenders
  http://www mdccorp com
  http://www engineers-international com/tenders html
  http://www canhealth.com/News040 html
  http://www lakemac com au/AboutCouncil/meetings asp
```

Figure 6.1.4b *CFT storage -test*

Both CFT storage and CFT sites accumulate CFT pertinent URLs each searching process. In figure 6.1.4b, the current searching city is Calgary, and the marked URL: http://www.ucalgary.ca. is added in the end of the CFT sites. At present, there are 829 saved URLs in CFT storage and 456 URLs in CFT sites. About 75% of the CFT storage or CFT sites are CFT pertinent URLs.

6.2.2 CFT URL

This part will evaluate the searching results of our system. At the same time input the same query term to Google. Comparing the top 10 responses with each other .

1) Top 10 of Domain specific searching system

CFT URLs are current searching results. There are 30 URL listed in the window. The first 10 are listed in the following page (see figure 6.1.4cb, there are more CFT URL tests, see appendix 2).

1) http://www.calgarywinterclub.com/constructionmar01.html

2) http://www.lexum.umontreal.ca/csc-cc/en/pub/1987/vol2/texte/1987scr2_0757.txt

3) http://www.gasandoil.com/contracts/samples/2004issue22.htm

4) http://surplus.gov.ab.ca/attachments/TNCALOUT_2003_044__High_River_.pdf

5) http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/vol2/html/1987scr2_0757.html

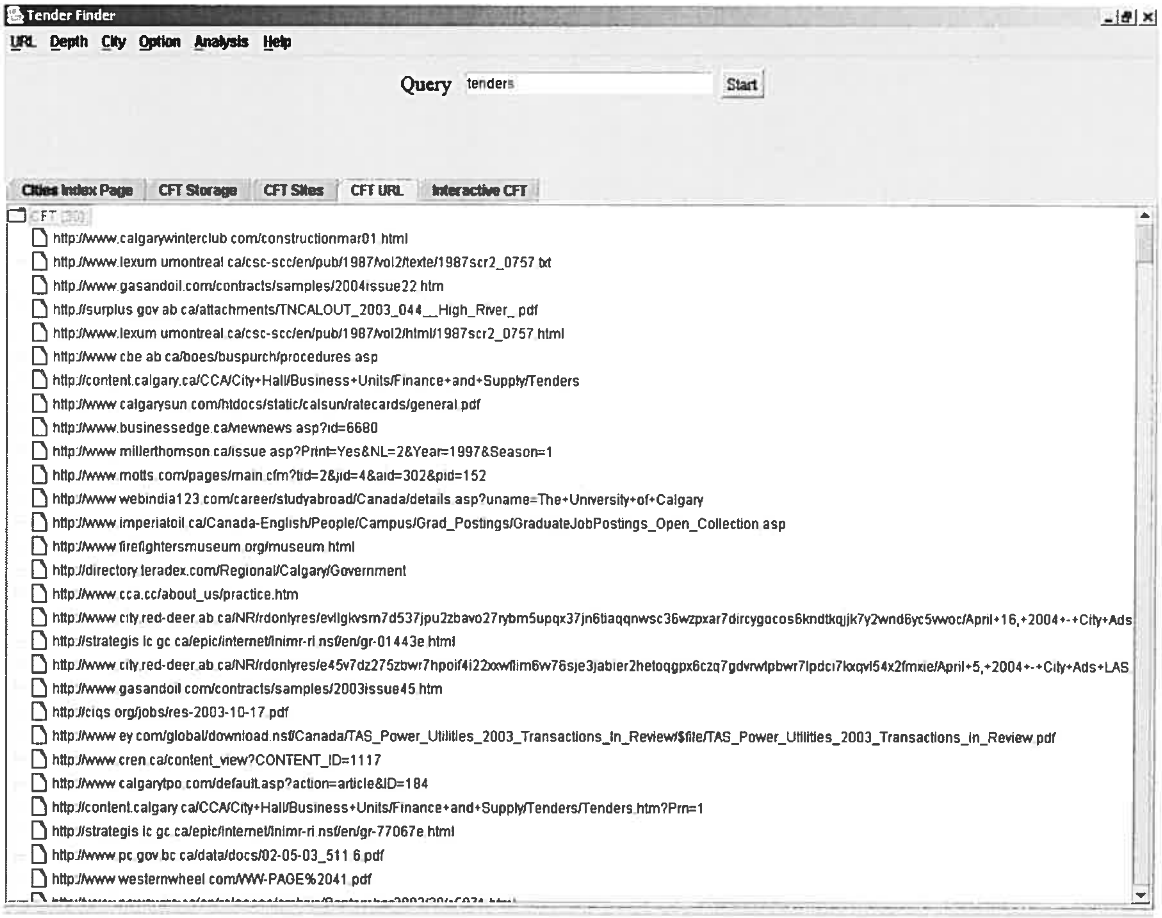
6) http://www.cbe.ab.ca/boes/buspurch/procedures.asp

7) http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders

8) http://www.calgarysun.com/htdocs/static/calsun/ratecards/general.pdf

9) http://www.businessedge.ca/viewnews.asp?id=6680

10) http://www.millerthomson.ca/issue.asp?Print=Yes&NL=2&Year=1997&Season=1

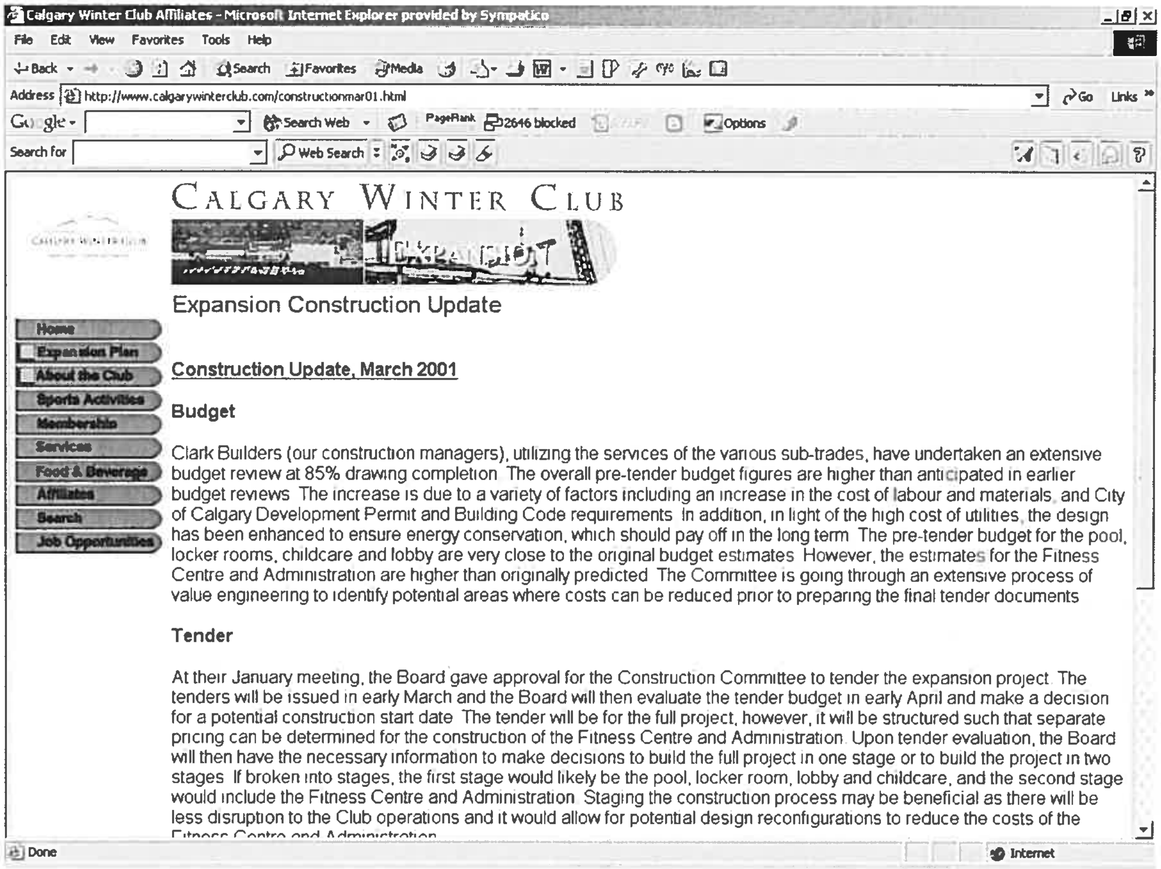


Query: tenders

```
CFT (30)
http://www.calgarywinterclub.com/constructionmar01.html
http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/vol2/texte/1987scr2_0757.txt
http://www.gasandoil.com/contracts/samples/2004issue22.htm
http://surplus.gov.ab.ca/attachments/TNCALOUT_2003_044__High_River_.pdf
http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/vol2/html/1987scr2_0757.html
http://www.cbe.ab.ca/boes/buspurch/procedures.asp
http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders
http://www.calgarysun.com/htdocs/static/calsun/ratecards/general.pdf
http://www.businessedge.ca/viewnews.asp?id=6680
http://www.millerthomson.ca/issue.asp?Print=Yes&NL=2&Year=1997&Season=1
http://www.motts.com/pages/main.cfm?tid=2&jid=4&aid=302&pid=152
http://www.webindia123.com/career/studyabroad/Canada/details.asp?uname=The+University+of+Calgary
http://www.imperialoil.ca/Canada-English/People/Campus/Grad_Postings/GraduateJobPostings_Open_Collection.asp
http://www.firefightersmuseum.org/museum.html
http://directory.teradex.com/Regional/Calgary/Government
http://www.cca.cc/about_us/practice.htm
http://www.city.red-deer.ab.ca/NR/rdonlyres/evilgkvsm7d537jpu2zbavo27rybm5upqx37jn6tiaqqnwsc36wzpxar7dircygocos6kndtkqjjk7y2wnd6yc5vwoc/April+16,+2004+-+City+Ads
http://strategis.ic.gc.ca/epic/internet/inimr-ri.nsf/en/gr-01443e.html
http://www.city.red-deer.ab.ca/NR/rdonlyres/e45v7dz275zbwr7hpoif4i22xxwflim6w76sje3jabier2hetoqgpx6czq7gdvrwtpbwr7lpdci7kxqvl54x2fmxie/April+5,+2004+-+City+Ads+LAS
http://www.gasandoil.com/contracts/samples/2003issue45.htm
http://ciqs.org/jobs/res-2003-10-17.pdf
http://www.ey.com/global/download.nsf/Canada/TAS_Power_Utilities_2003_Transactions_In_Review/$file/TAS_Power_Utilities_2003_Transactions_In_Review.pdf
http://www.cren.ca/content_view?CONTENT_ID=1117
http://www.calgarytpo.com/default.asp?action=article&ID=184
http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders/Tenders.htm?Prn=1
http://strategis.ic.gc.ca/epic/internet/inimr-ri.nsf/en/gr-77067e.html
http://www.pc.gov.bc.ca/data/docs/02-05-03_511.6.pdf
http://www.westernwheel.com/WW-PAGE%2041.pdf
```

Figure 6.1.4c *CFT URLs –test*

Let's evaluate this system's response. Figure 6.1.5a shows the content of http://www.calgarywinterclub.com/constructionmar01.html


Calgary Winter Club Affiliates - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.calgarywinterclub.com/constructionmar01.html
Go Links
Google Search Web PageRank 2646 blocked Options
Search for Web Search
Home
Expansion Plan
About the Club
Sports Activities
Membership
Services
Food & Beverage
Affiliates
Search
Job Opportunities
Done
Internet

CALGARY WINTER CLUB

Expansion Construction Update

Construction Update, March 2001

Budget

Clark Builders (our construction managers), utilizing the services of the various sub-trades, have undertaken an extensive budget review at 85% drawing completion The overall pre-tender budget figures are higher than anticipated in earlier budget reviews The increase is due to a variety of factors including an increase in the cost of labour and materials, and City of Calgary Development Permit and Building Code requirements In addition, in light of the high cost of utilities, the design has been enhanced to ensure energy conservation, which should pay off in the long term The pre-tender budget for the pool, locker rooms, childcare and lobby are very close to the original budget estimates However, the estimates for the Fitness Centre and Administration are higher than originally predicted The Committee is going through an extensive process of value engineering to identify potential areas where costs can be reduced prior to preparing the final tender documents

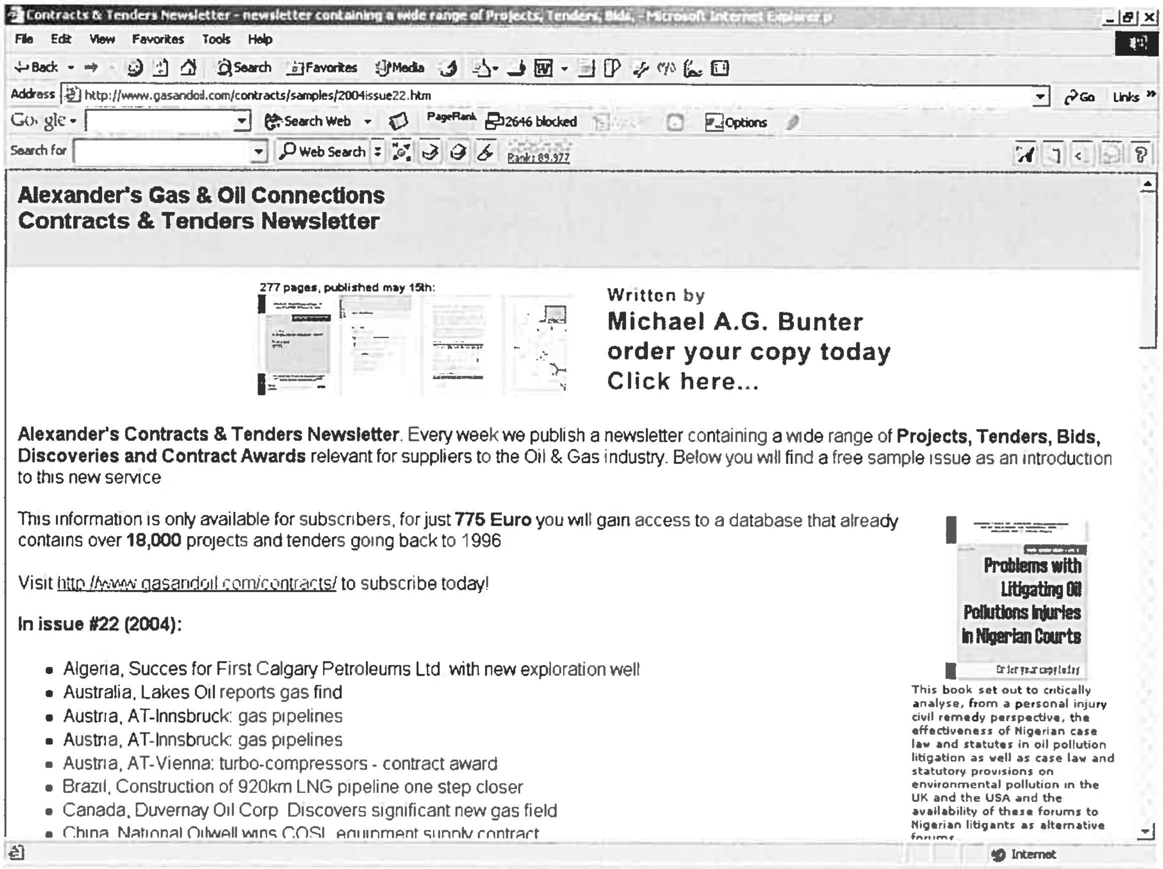
Tender

At their January meeting, the Board gave approval for the Construction Committee to tender the expansion project. The tenders will be issued in early March and the Board will then evaluate the tender budget in early April and make a decision for a potential construction start date The tender will be for the full project, however, it will be structured such that separate pricing can be determined for the construction of the Fitness Centre and Administration. Upon tender evaluation, the Board will then have the necessary information to make decisions to build the full project in one stage or to build the project in two stages If broken into stages, the first stage would likely be the pool, locker room, lobby and childcare, and the second stage would include the Fitness Centre and Administration Staging the construction process may be beneficial as there will be less disruption to the Club operations and it would allow for potential design reconfigurations to reduce the costs of the Fitness Centre and Administration

Figure 6.1.5a *CFT URLs –response evaluation*

2) http://www.lexum.umontreal.ca/csc-cc/en/pub/1987/vol2/texte/1987scr2_0757.txt file is not found at testing moment;

3) http://www.gasandoil.com/contracts/samples/2004issue22.htm is a good URL point to a CFT in Calgary.


Contracts & Tenders Newsletter - newsletter containing a wide range of Projects, Tenders, Bids, - Microsoft Internet Explorer p
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.gasandoil.com/contracts/samples/2004issue22.htm
Go Links
Google Search Web PageRank 2646 blocked Options
Search for Web Search
Internet

Alexander's Gas & Oil Connections
Contracts & Tenders Newsletter

277 pages, published may 15th:

Written by
Michael A.G. Bunter
order your copy today
Click here...

Alexander's Contracts & Tenders Newsletter. Every week we publish a newsletter containing a wide range of **Projects, Tenders, Bids, Discoveries and Contract Awards** relevant for suppliers to the Oil & Gas industry. Below you will find a free sample issue as an introduction to this new service

This information is only available for subscribers, for just **775 Euro** you will gain access to a database that already contains over **18,000** projects and tenders going back to 1996

Visit http://www.gasandoil.com/contracts/ to subscribe today!

In issue #22 (2004):

- Algeria, Succes for First Calgary Petroleums Ltd with new exploration well
- Australia, Lakes Oil reports gas find
- Austria, AT-Innsbruck: gas pipelines
- Austria, AT-Innsbruck: gas pipelines
- Austria, AT-Vienna: turbo-compressors - contract award
- Brazil, Construction of 920km LNG pipeline one step closer
- Canada, Duvernay Oil Corp Discovers significant new gas field
- China, National Oilwell wins COSL equipment supply contract

Problems with Litigating Oil Pollutions Injuries in Nigerian Courts

This book set out to critically analyse, from a personal injury civil remedy perspective, the effectiveness of Nigerian case law and statutes in oil pollution litigation as well as case law and statutory provisions on environmental pollution in the UK and the USA and the availability of these forums to Nigerian litigants as alternative forums

Figure 6.1.5b *CFT URLs –response evaluation*

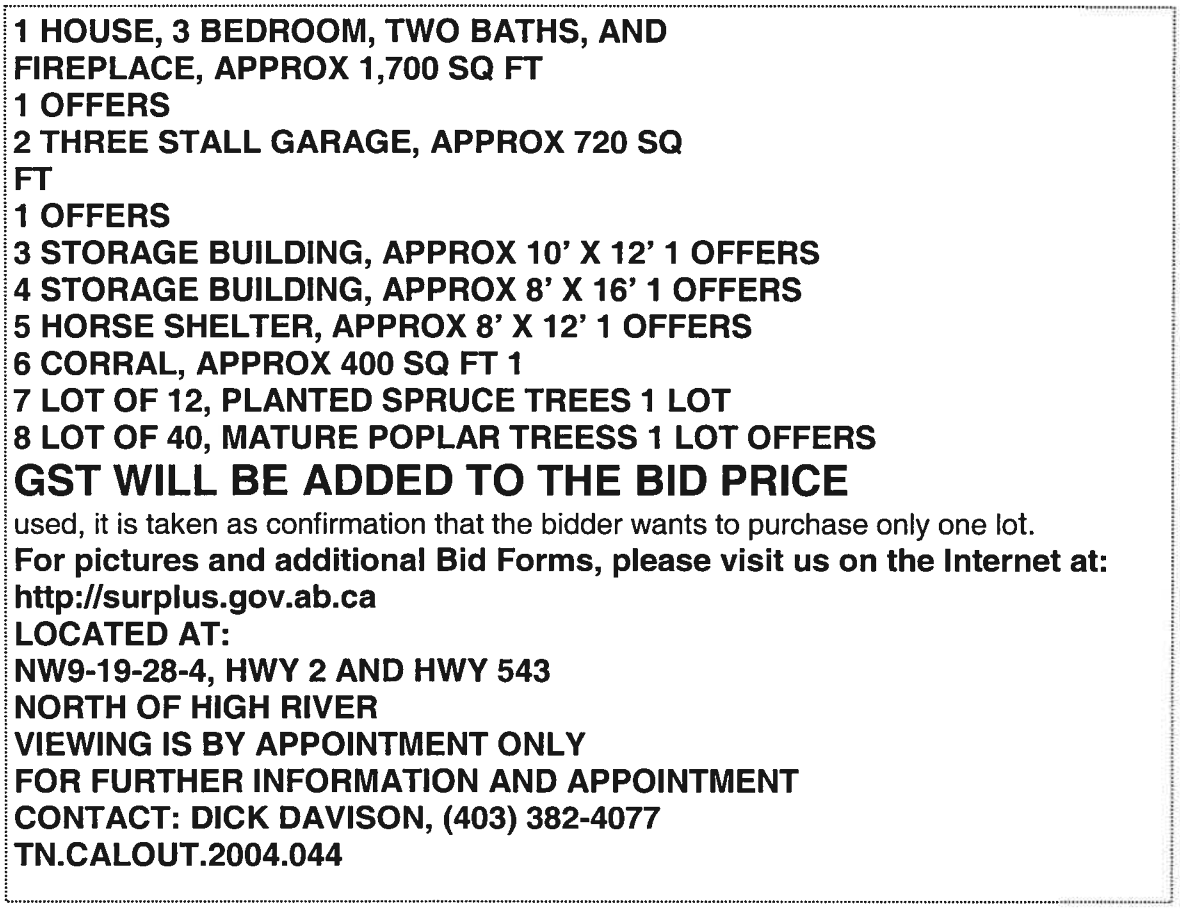
4) http://surplus.gov.ab.ca/attachments/TNCALOUT_2003_044__High_River_.pdf is a PDF file. The content is a CFT document in Calgary. See the PDF content in figure 6.1.5c in the following.

A **SURPLUS SALES CALGARY**
CORPORATE SERVICE CENTRE TN.CALOUT.2004.044
3850 MANCHESTER RD, SE
CALGARY, AB. T2G 3Z9
PHONE: (403) 297-6430
FAX: (403) 297-4576

INVITATION TO TENDER

TENDER No:

Sealed Tenders, subject to the General Terms and Conditions – Public Tenders and any Special Terms and Conditions,
will be received at the office of Alberta Corporate Service Centre, Surplus Sales Section at the above address, **NOT**
LATER THAN: 14:00:59 Alberta Time, Tuesday, September 21, 2004
Tenders will be opened in public at the above noted address, time and date, FOR THE SALE OF:
Lot No. Description Quantity



1 HOUSE, 3 BEDROOM, TWO BATHS, AND
FIREPLACE, APPROX 1,700 SQ FT
1 OFFERS
2 THREE STALL GARAGE, APPROX 720 SQ
FT
1 OFFERS
3 STORAGE BUILDING, APPROX 10' X 12' 1 OFFERS
4 STORAGE BUILDING, APPROX 8' X 16' 1 OFFERS
5 HORSE SHELTER, APPROX 8' X 12' 1 OFFERS
6 CORRAL, APPROX 400 SQ FT 1
7 LOT OF 12, PLANTED SPRUCE TREES 1 LOT
8 LOT OF 40, MATURE POPLAR TREESS 1 LOT OFFERS

GST WILL BE ADDED TO THE BID PRICE

used, it is taken as confirmation that the bidder wants to purchase only one lot.

For pictures and additional Bid Forms, please visit us on the Internet at:
http://surplus.gov.ab.ca

LOCATED AT:
NW9-19-28-4, HWY 2 AND HWY 543
NORTH OF HIGH RIVER
VIEWING IS BY APPOINTMENT ONLY
FOR FURTHER INFORMATION AND APPOINTMENT
CONTACT: DICK DAVISON, (403) 382-4077
TN.CALOUT.2004.044

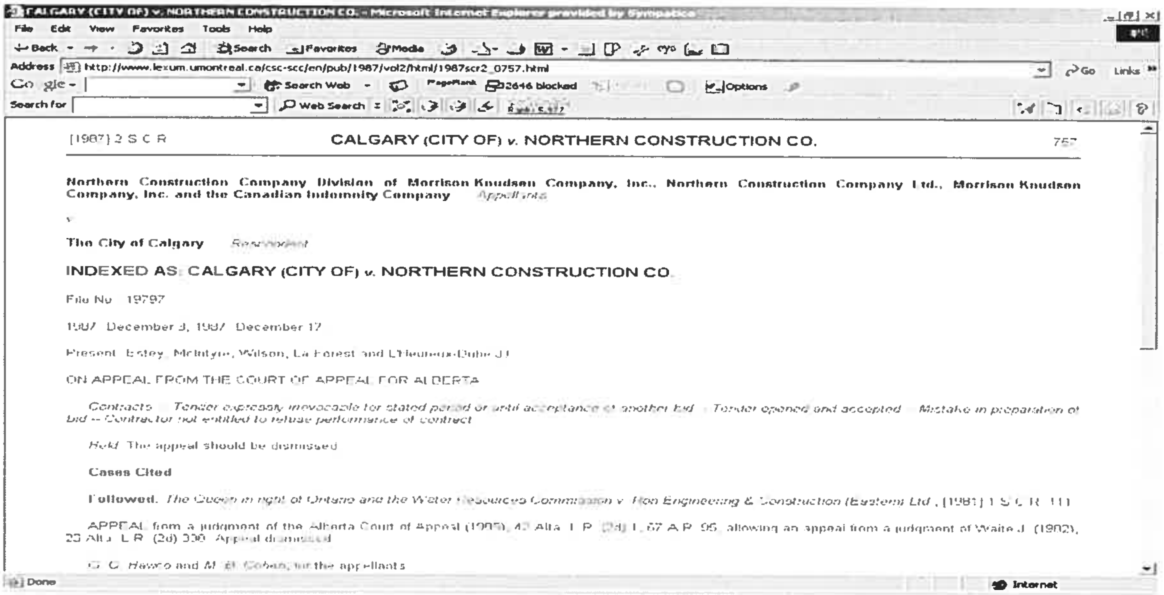
Figure 6.1.5c *CFT URLs – response evaluation*

5) http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/vol2/html/1987scr2_0757.html is tender related page, but strictly speaking, not a CFT (Figure 6.1.5d).

6) http://www.cbe.ab.ca/boes/buspurch/procedures.asp is a CFT in Calgary (Figure 6.1.5e).

7) is a tender hub where indicate to *Commodities Tenders, Construction Tenders and Services Tenders* (see four Figure 6.1.5f - Figure 6.1.5f (3)).

http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders


CALGARY (CITY OF) v. NORTHERN CONSTRUCTION CO. - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Go Links
Search Web PageRank 2646 blocked Options
Search for Web Search
Done
Internet

Address http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/vol2/html/1987scr2_0757.html

[1987] 2 S C R CALGARY (CITY OF) v. NORTHERN CONSTRUCTION CO. 757

Northern Construction Company Division of Morrison Knudsen Company, Inc., Northern Construction Company Ltd., Morrison Knudsen Company, Inc. and the Canadian Indemnity Company *Appellants*

v

The City of Calgary *Respondent*

INDEXED AS: CALGARY (CITY OF) v. NORTHERN CONSTRUCTION CO.

File No 19797

1987 December 3, 1987 December 17

Present Estey, McIntyre, Wilson, La Forest and L'Heureux-Dube JJ

ON APPEAL FROM THE COURT OF APPEAL FOR ALBERTA

Contracts — Tender expressly irrevocable for stated period or until acceptance of another bid — Tender opened and accepted — Mistake in preparation of bid -- Contractor not entitled to refuse performance of contract

Held The appeal should be dismissed

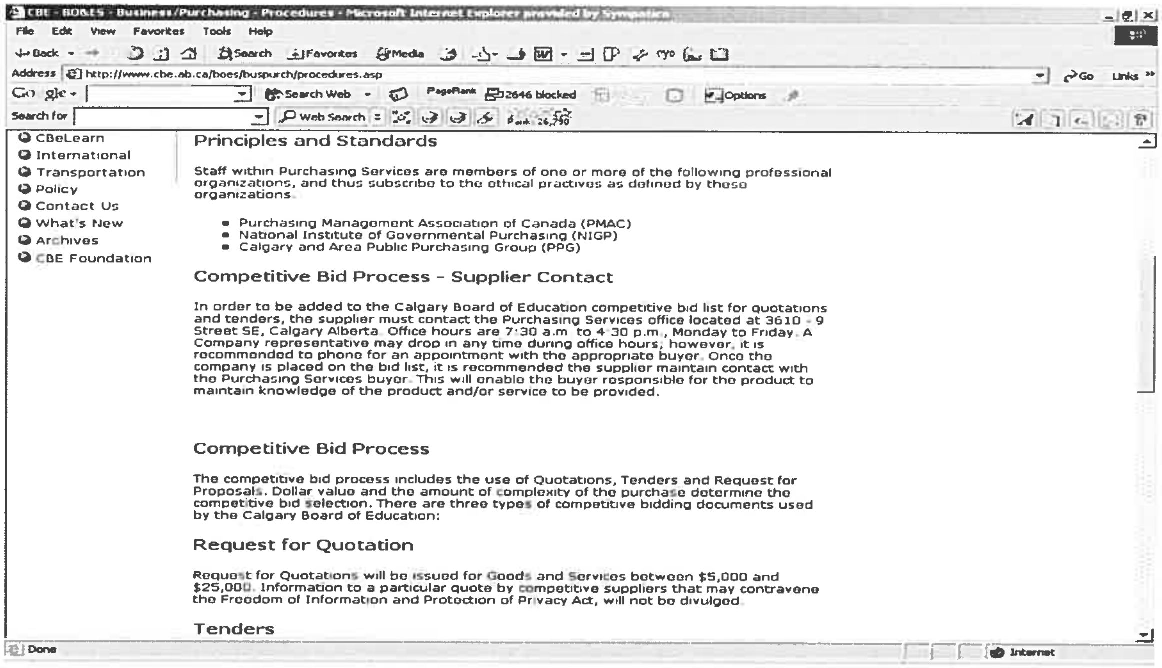
Cases Cited

Followed. *The Queen in right of Ontario and the Water Resources Commission v. Ron Engineering & Construction (Eastern) Ltd*, [1981] 1 S C R 111

APPEAL from a judgment of the Alberta Court of Appeal (1985), 42 Alta L R (2d) 1, 67 A R 95, allowing an appeal from a judgment of Waite J (1982), 23 Alta L R (2d) 338 Appeal dismissed

G C *Hawco* and *M El Cohen*, for the appellants

Figure 6.1.5d *CFT URLs – response evaluation*


CBE - BOBES - Business/Purchasing - Procedures - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Go Links
Search Web PageRank 2646 blocked Options
Search for Web Search
CBeLearn
International
Transportation
Policy
Contact Us
What's New
Archives
CBE Foundation
Done
Internet

Address http://www.cbe.ab.ca/boes/buspurch/procedures.asp

Principles and Standards

Staff within Purchasing Services are members of one or more of the following professional organizations, and thus subscribe to the ethical practices as defined by those organizations.

- Purchasing Management Association of Canada (PMAC)
- National Institute of Governmental Purchasing (NIGP)
- Calgary and Area Public Purchasing Group (PPG)

Competitive Bid Process - Supplier Contact

In order to be added to the Calgary Board of Education competitive bid list for quotations and tenders, the supplier must contact the Purchasing Services office located at 3610 - 9 Street SE, Calgary Alberta. Office hours are 7:30 a.m. to 4:30 p.m., Monday to Friday. A Company representative may drop in any time during office hours, however, it is recommended to phone for an appointment with the appropriate buyer. Once the company is placed on the bid list, it is recommended the supplier maintain contact with the Purchasing Services buyer. This will enable the buyer responsible for the product to maintain knowledge of the product and/or service to be provided.

Competitive Bid Process

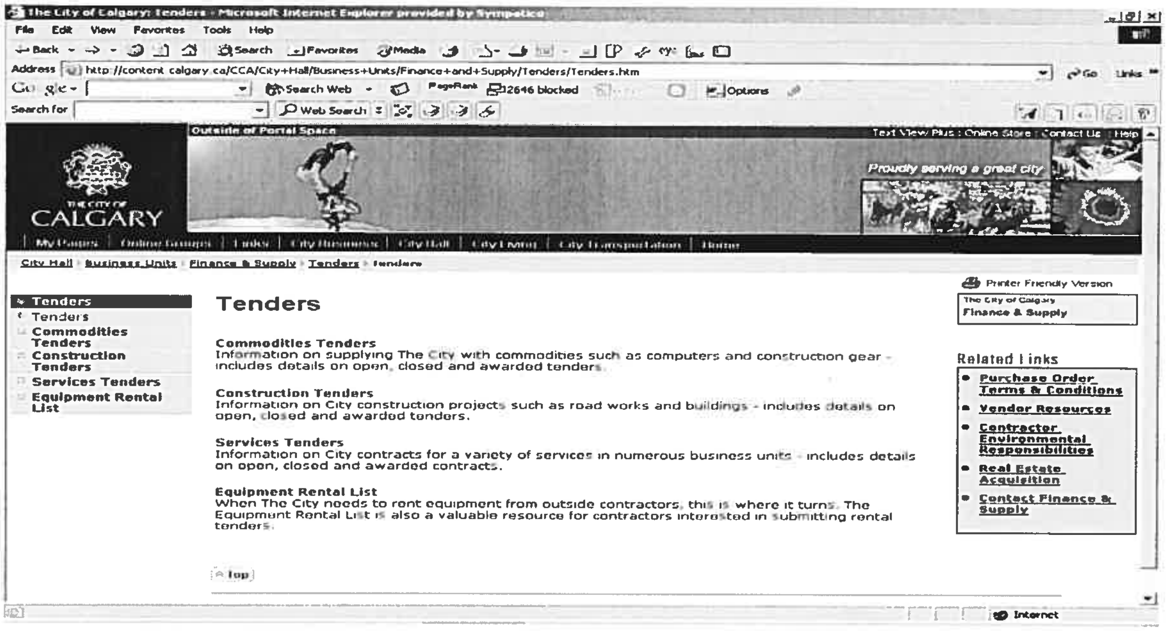
The competitive bid process includes the use of Quotations, Tenders and Request for Proposals. Dollar value and the amount of complexity of the purchase determine the competitive bid selection. There are three types of competitive bidding documents used by the Calgary Board of Education:

Request for Quotation

Request for Quotations will be issued for Goods and Services between $5,000 and $25,000. Information to a particular quote by competitive suppliers that may contravene the Freedom of Information and Protection of Privacy Act, will not be divulged.

Tenders

Figure 6.1.5e *CFT URLs – response evaluation*


The City of Calgary: Tenders - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://content calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders/Tenders.htm
Go Links
Search Web PageRank 2646 blocked Options
Search for Web Search
Outside of Portal Space
Text View Plus : Online Store : Contact Us : Help
THE CITY OF CALGARY
Proudly serving a great city
City Hall › Business Units › Finance & Supply › Tenders › Tenders
Printer Friendly Version
The City of Calgary Finance & Supply
Tenders
Tenders
Commodities Tenders
Construction Tenders
Services Tenders
Equipment Rental List
top
Internet

Tenders

Commodities Tenders
Information on supplying The City with commodities such as computers and construction gear - includes details on open, closed and awarded tenders.

Construction Tenders
Information on City construction projects such as road works and buildings - includes details on open, closed and awarded tenders.

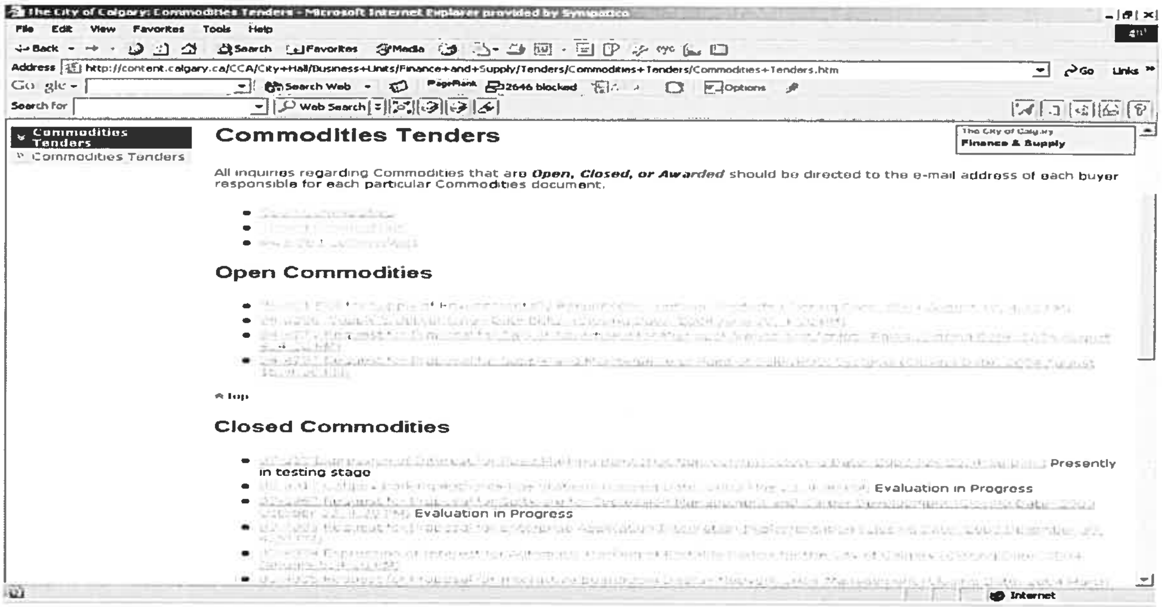
Services Tenders
Information on City contracts for a variety of services in numerous business units - includes details on open, closed and awarded contracts.

Equipment Rental List
When The City needs to rent equipment from outside contractors, this is where it turns. The Equipment Rental List is also a valuable resource for contractors interested in submitting rental tenders.

Related Links

- Purchase Order Terms & Conditions
- Vendor Resources
- Contractor Environmental Responsibilities
- Real Estate Acquisition
- Contact Finance & Supply

Figure 6.1.5f *CFT URLs – response evaluation*


The City of Calgary: Commodities Tenders - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders/Commodities+Tenders/Commodities+Tenders.htm
Go Links
Search Web PageRank 2646 blocked Options
Search for Web Search
Commodities Tenders
Commodities Tenders
The City of Calgary Finance & Supply
top
Internet

Commodities Tenders

All inquiries regarding Commodities that are ***Open, Closed, or Awarded*** should be directed to the e-mail address of each buyer responsible for each particular Commodities document.

- [illegible]
- [illegible]
- [illegible]

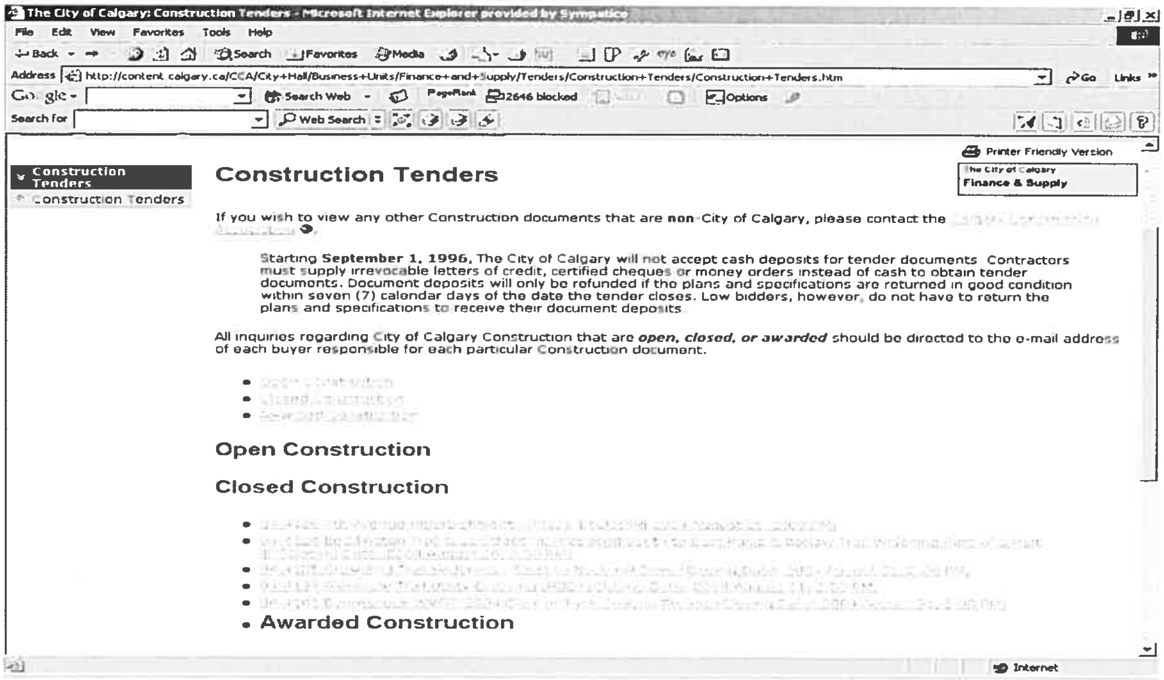
Open Commodities

- [illegible]
- [illegible]
- [illegible]
- [illegible]

Closed Commodities

- [illegible] Presently in testing stage
- [illegible] Evaluation in Progress
- [illegible] Evaluation in Progress
- [illegible]
- [illegible]
- [illegible]

Figure 6.1.5f (1) *Commodities tenders*


The City of Calgary: Construction Tenders - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders/Construction+Tenders/Construction+Tenders.htm
Go Links
Search Web PageRank 2646 blocked Options
Search for Web Search
Printer Friendly Version
The City of Calgary Finance & Supply
Construction Tenders
Construction Tenders
Internet

Construction Tenders

If you wish to view any other Construction documents that are **non**-City of Calgary, please contact the [illegible]

> Starting **September 1, 1996**, The City of Calgary will not accept cash deposits for tender documents. Contractors must supply irrevocable letters of credit, certified cheques or money orders instead of cash to obtain tender documents. Document deposits will only be refunded if the plans and specifications are returned in good condition within seven (7) calendar days of the date the tender closes. Low bidders, however, do not have to return the plans and specifications to receive their document deposits.

All inquiries regarding City of Calgary Construction that are ***open, closed, or awarded*** should be directed to the e-mail address of each buyer responsible for each particular Construction document.

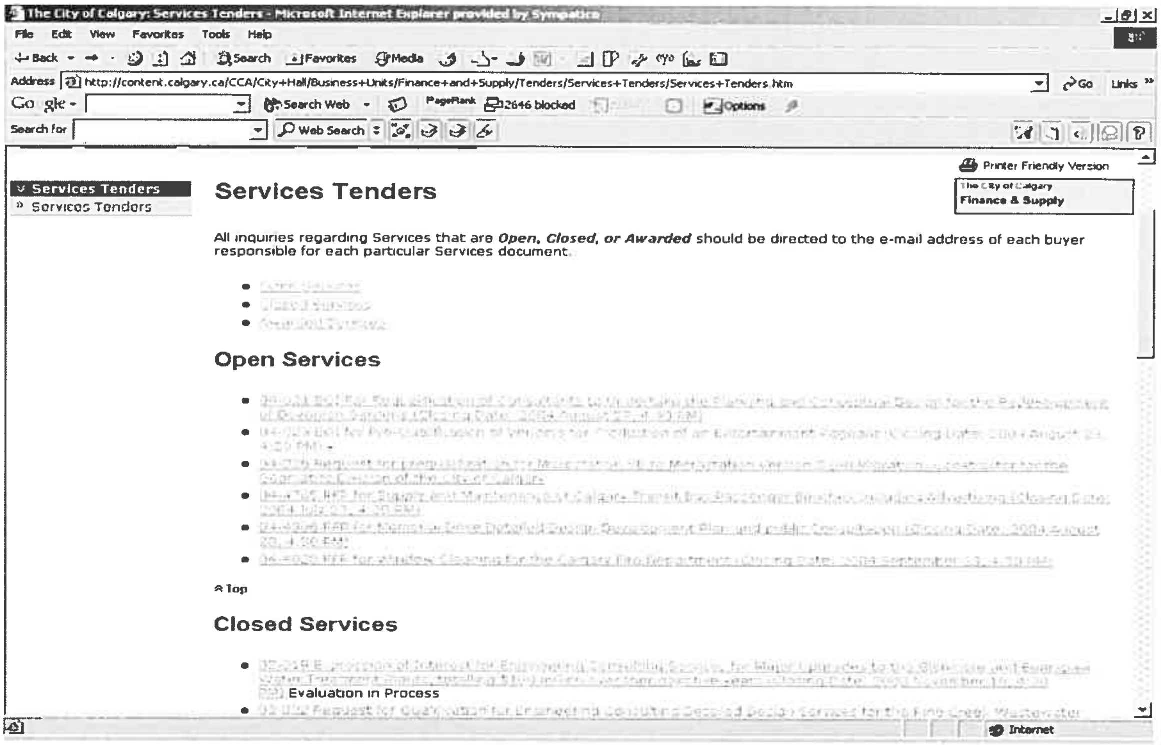
- [illegible]
- [illegible]
- [illegible]

Open Construction

Closed Construction

- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- **Awarded Construction**

Figure 6.1.5f (2) *Construction tenders*


The City of Calgary: Services Tenders - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders/Services+Tenders/Services+Tenders.htm
Go Links
Search Web PageRank 2646 blocked Options
Search for Web Search
Printer Friendly Version
The City of Calgary Finance & Supply
Services Tenders
» Services Tenders
Internet

Services Tenders

All inquiries regarding Services that are ***Open, Closed, or Awarded*** should be directed to the e-mail address of each buyer responsible for each particular Services document.

- [illegible]
- [illegible]
- [illegible]

Open Services

- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]

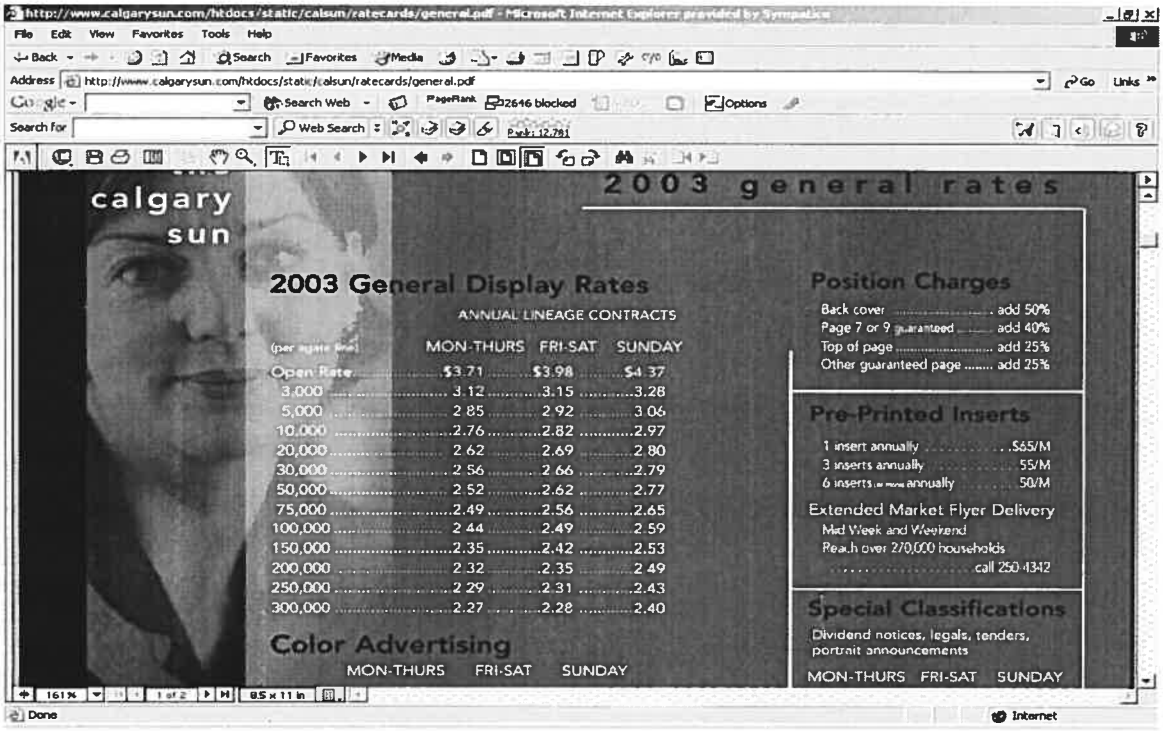
^ Top

Closed Services

- [illegible] Evaluation in Process
- [illegible]

Figure 6.1.5f (3) *Services tenders*

8) http://www.calgarysun.com/htdocs/static/calsun/ratecards/general.pdf is just a advertisement of a newspaper. Similar to a tender of advertisement, but not a real CFT (Figure 6.1.5g).


http://www.calgarysun.com/htdocs/static/calsun/ratecards/general.pdf - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.calgarysun.com/htdocs/static/calsun/ratecards/general.pdf
Go Links
Search Web PageRank 2646 blocked Options
Search for Web Search
161% 1 of 2 8.5 x 11 in
Done
Internet
calgary sun

2003 general rates

2003 General Display Rates

ANNUAL LINEAGE CONTRACTS

(per agate line)	MON-THURS	FRI-SAT	SUNDAY
Open Rate	$3.71	$3.98	$4.37
3,000	3.12	3.15	3.28
5,000	2.85	2.92	3.06
10,000	2.76	2.82	2.97
20,000	2.62	2.69	2.80
30,000	2.56	2.66	2.79
50,000	2.52	2.62	2.77
75,000	2.49	2.56	2.65
100,000	2.44	2.49	2.59
150,000	2.35	2.42	2.53
200,000	2.32	2.35	2.49
250,000	2.29	2.31	2.43
300,000	2.27	2.28	2.40

Color Advertising

MON-THURS FRI-SAT SUNDAY

Position Charges

Back cover	add 50%
Page 7 or 9 guaranteed	add 40%
Top of page	add 25%
Other guaranteed page	add 25%

Pre-Printed Inserts

1 insert annually	$65/M
3 inserts annually	55/M
6 inserts or more annually	50/M

Extended Market Flyer Delivery
Mid Week and Weekend
Reach over 270,000 households
call 250-4342

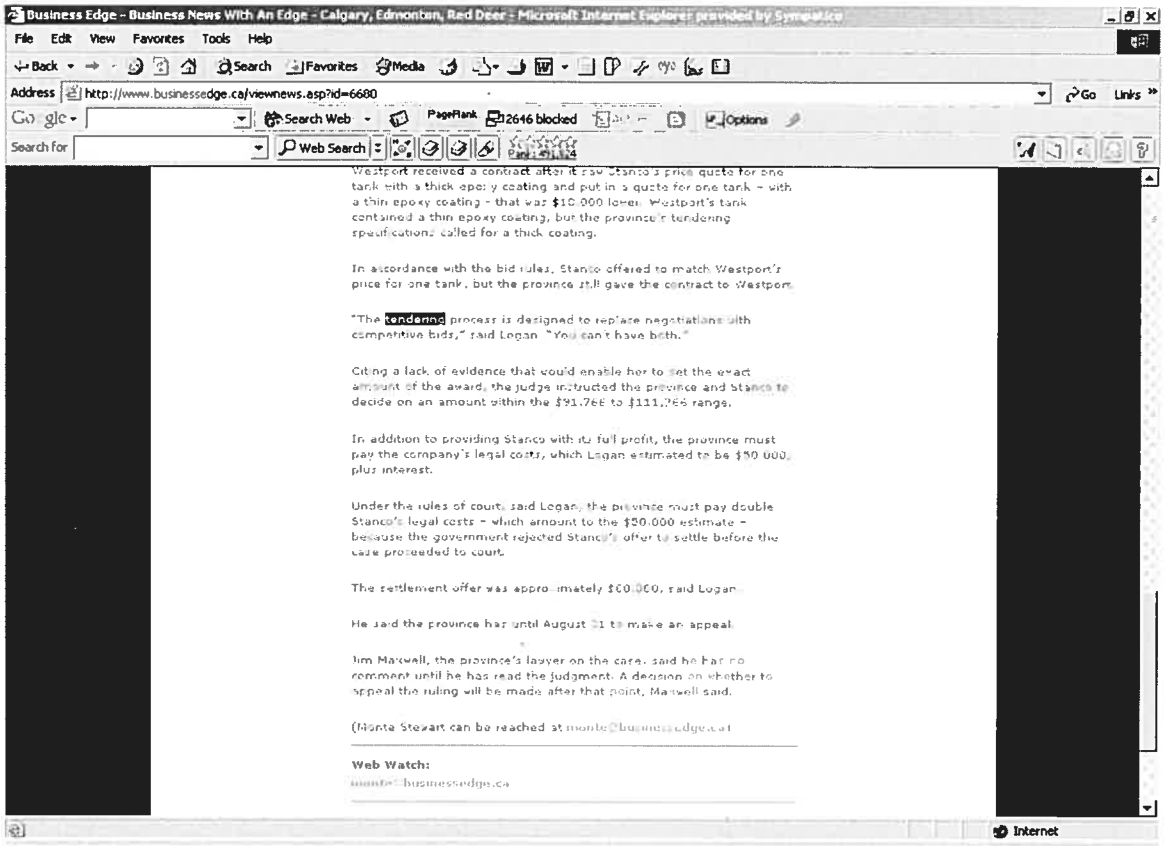
Special Classifications

Dividend notices, legals, tenders, portrait announcements

MON-THURS FRI-SAT SUNDAY

Figure 6.1.5g *Services tenders*

9) http://www.businessedge.ca/viewnews.asp?id=6680 is shown in Figure 6.1.5h in the next page. It's a tender pertinent report which published at 8/19/2004 –Vol. 1, No.17. The title is "**Bid shopping' accusation stings province**", and sub-title is "**Contractors applaud ruling that forbids hybrid process By Monte Stewart - Business Edge**". Actually this is a typical example of CFT confusion type. It is really a difficulty one for crawler to distinguish this page as a CFT or not. Because from lexical view, it meet nearly all the need of CFT, such as tender, value, date, time etc. we'll discuss this type of page in next chapter.


Business Edge - Business News With An Edge - Calgary, Edmonton, Red Deer - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.businessedge.ca/viewnews.asp?id=6680
Go Links
Google Search Web PageRank 2646 blocked Options
Search for Web Search
Internet

Westport received a contract after it saw Stanco's price quote for one tank with a thick epoxy coating and put in a quote for one tank – with a thin epoxy coating – that was $10,000 lower. Westport's tank contained a thin epoxy coating, but the province's tendering specifications called for a thick coating.

In accordance with the bid rules, Stanco offered to match Westport's price for one tank, but the province still gave the contract to Westport.

"The tendering process is designed to replace negotiations with competitive bids," said Logan. "You can't have both."

Citing a lack of evidence that would enable her to set the exact amount of the award, the judge instructed the province and Stanco to decide on an amount within the $91,766 to $111,766 range.

In addition to providing Stanco with its full profit, the province must pay the company's legal costs, which Logan estimated to be $50,000, plus interest.

Under the rules of court, said Logan, the province must pay double Stanco's legal costs – which amount to the $50,000 estimate – because the government rejected Stanco's offer to settle before the case proceeded to court.

The settlement offer was approximately $60,000, said Logan.

He said the province has until August 31 to make an appeal.

Jim Maxwell, the province's lawyer on the case, said he has no comment until he has read the judgment. A decision on whether to appeal the ruling will be made after that point, Maxwell said.

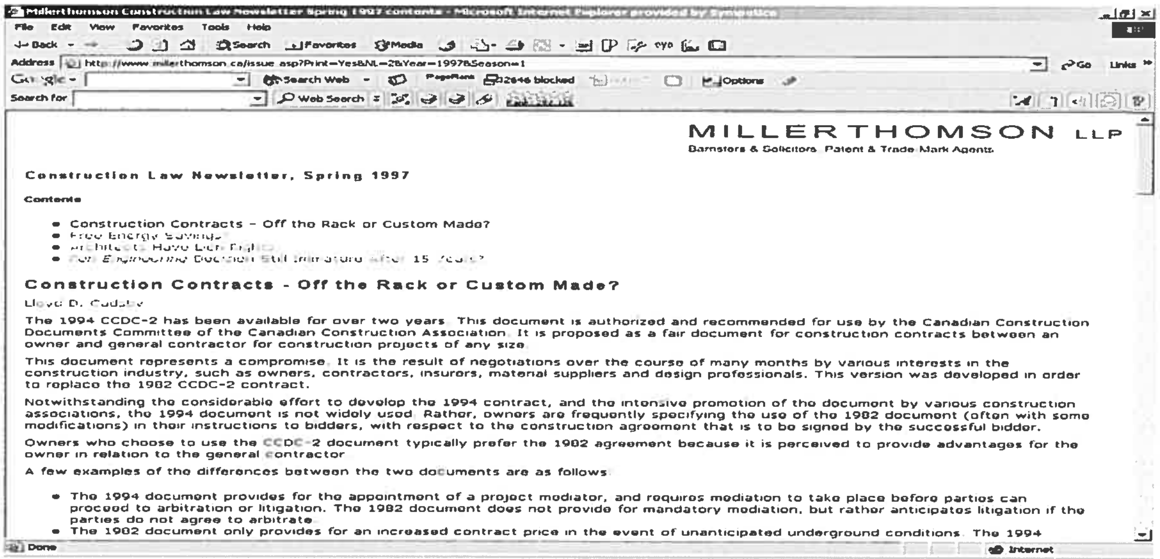
(Monte Stewart can be reached at monte@businessedge.ca)

Web Watch:

monte@businessedge.ca

Figure 6.1.5h *CFT URLs – response evaluation*

10) http://www.millerthomson.ca/issue.asp?Print=Yes&NL=2&Year=1997&Season=1 is an article talking about construction law which pertinent with the construction tender.


Miller Thomson Construction Law Newsletter Spring 1997 contents - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.millerthomson.ca/issue.asp?Print=Yes&NL=2&Year=1997&Season=1
Go Links
Google Search Web PageRank 2646 blocked Options
Search for Web Search
Done
Internet

MILLER THOMSON LLP
Barristers & Solicitors Patent & Trade Mark Agents

Construction Law Newsletter, Spring 1997

Contents

- Construction Contracts – Off the Rack or Custom Made?
- Free Energy Savings
- Architects Have Lien Rights
- Fort Engineering Decision Still Immature After 15 Years?

Construction Contracts - Off the Rack or Custom Made?

Lloyd D. Cadsby

The 1994 CCDC-2 has been available for over two years. This document is authorized and recommended for use by the Canadian Construction Documents Committee of the Canadian Construction Association. It is proposed as a fair document for construction contracts between an owner and general contractor for construction projects of any size.

This document represents a compromise. It is the result of negotiations over the course of many months by various interests in the construction industry, such as owners, contractors, insurers, material suppliers and design professionals. This version was developed in order to replace the 1982 CCDC-2 contract.

Notwithstanding the considerable effort to develop the 1994 contract, and the intensive promotion of the document by various construction associations, the 1994 document is not widely used. Rather, owners are frequently specifying the use of the 1982 document (often with some modifications) in their instructions to bidders, with respect to the construction agreement that is to be signed by the successful bidder.

Owners who choose to use the CCDC-2 document typically prefer the 1982 agreement because it is perceived to provide advantages for the owner in relation to the general contractor.

A few examples of the differences between the two documents are as follows:

- The 1994 document provides for the appointment of a project mediator, and requires mediation to take place before parties can proceed to arbitration or litigation. The 1982 document does not provide for mandatory mediation, but rather anticipates litigation if the parties do not agree to arbitrate.
- The 1982 document only provides for an increased contract price in the event of unanticipated underground conditions. The 1994

Figure 6.1.5i *CFT URLs – response evaluation*

The title of the 10) pointed article is "Construction Contracts - Off the Rack or Custom Made?"

Summarizing above top 10 searching results gets following table 6.1.1.

Category	Quantity	Percentage	Comment
CFT in Calgary	6	60%	1),3),4),5),6),7)
Similar to CFT	1	10%	8); in Calgary
File not Found	1	10%	2)
Others	2	20%	9) CFT news;10)Construction law

Table 6.1.1 *Summarizing of evaluation*

2) Top 10 of Google

At the same time, inputting the "Calgary tenders", top 10 search results provided by Google are :

a) http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders/Tenders.htm

b) www.westernwheel.com/WW-PAGE%2041.pdf

c) surplus.gov.ab.ca/attachments/ TNCALOUT_2003_044__High_River_.pdf

d) www.lexum.umontreal.ca/csc-scc/ en/pub/1987/vol2/html/1987scr2_0757.html

e) www.lexum.umontreal.ca/csc-scc/ en/pub/1987/vol2/texte/1987scr2_0757.txt

f) www.cca.cc/about_us/practice.htm

g) www.cbe.ab.ca/boes/buspurch/procedures.asp

h) directory.teradex.com/Regional/Calgary/Government

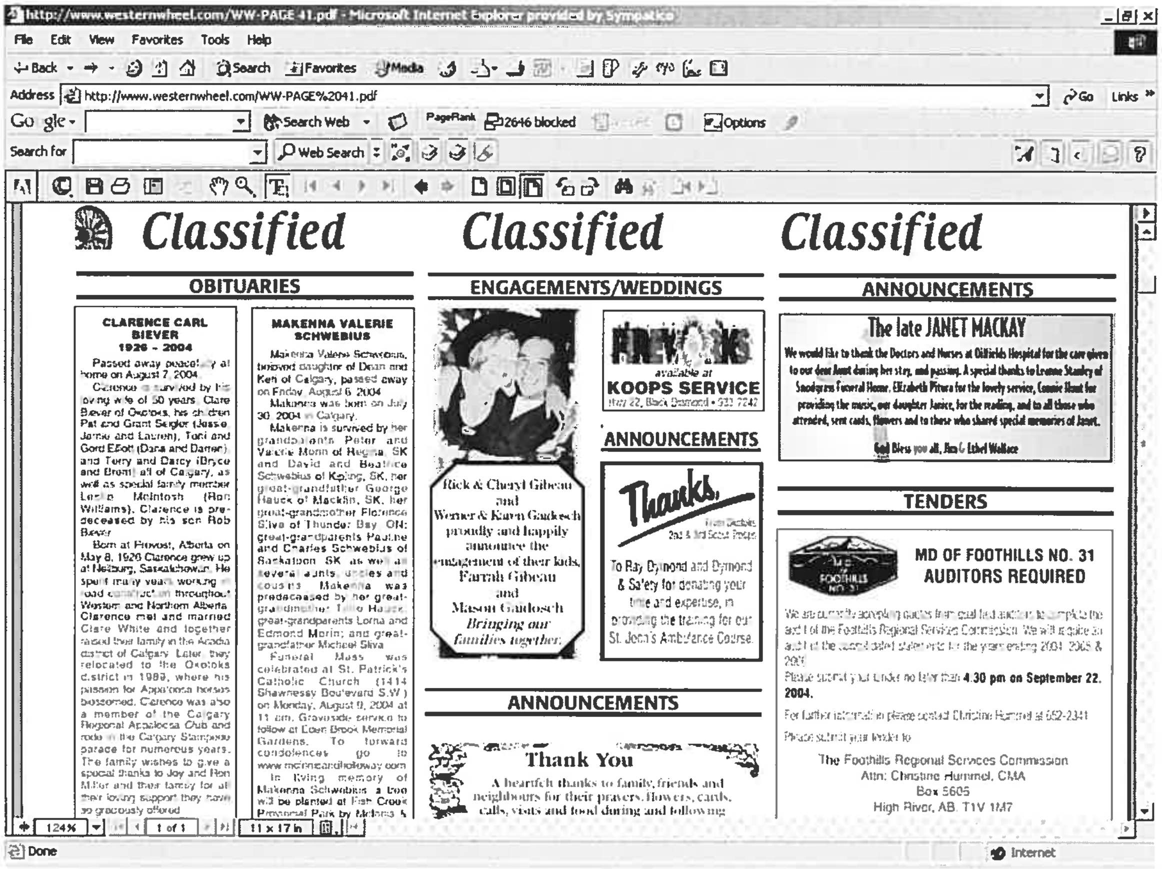
i) ciqs.org/jobs/res-2003-10-17.pdf

j) www.pc.gov.bc.ca/data/docs/02-05-03_511.6.pdf

Let's evaluating Google's top 10 response results one by one:

a) same as 7) a CFT hub point to CFTs in Calgary;

b) see the figure 6.1.6a, a tender advertisement nested in a newspaper.


http://www.westernwheel.com/WW-PAGE 41.pdf - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.westernwheel.com/WW-PAGE%2041.pdf
Go Links
Google Search Web PageRank 2646 blocked Options
Search for Web Search
124% 1 of 1 11 x 17 in
Done
Internet

Classified Classified Classified

OBITUARIES

CLARENCE CARL BIEVER
1926 - 2004

Passed away peacefully at home on August 7, 2004.

Clarence is survived by his loving wife of 50 years, Clare Biever of Okotoks, his children Pat and Grant Seigler (Jessie, Jamie and Lauren), Toni and Gord Elliott (Dana and Darren), and Terry and Darcy (Bryce and Brent) all of Calgary, as well as special family member Leslie McIntosh (Ron Williams). Clarence is predeceased by his son Rob Biever.

Born at Provost, Alberta on May 8, 1926 Clarence grew up at Neilburg, Saskatchewan. He spent many years working in road construction throughout Western and Northern Alberta. Clarence met and married Clare White and together raised their family in the Acadia district of Calgary. Later, they relocated to the Okotoks district in 1989, where his passion for Appaloosa horses blossomed. Clarence was also a member of the Calgary Regional Appaloosa Club and rode in the Calgary Stampede parade for numerous years. The family wishes to give a special thanks to Joy and Ron Miller and their family for all their loving support they have so graciously offered

MAKENNA VALERIE SCHWEBIUS

Makenna Valerie Schwebius, beloved daughter of Dean and Keri of Calgary, passed away on Friday, August 6, 2004.

Makenna was born on July 30, 2004 in Calgary.

Makenna is survived by her grandparents Peter and Valerie Morin of Regina, SK and David and Beatrice Schwebius of Kipling, SK, her great-grandfather George Hauck of Macklin, SK, her great-grandmother Florence Silva of Thunder Bay, ON; great-grandparents Pauline and Charles Schwebius of Saskatoon, SK as well as several aunts, uncles and cousins. Makenna was predeceased by her great-grandmother Tillie Hauck; great-grandparents Lorna and Edmond Morin; and great-grandfather Michael Silva.

Funeral Mass was celebrated at St. Patrick's Catholic Church (1414 Shawnessy Boulevard S.W.) on Monday, August 9, 2004 at 11 am. Graveside service to follow at Eden Brook Memorial Gardens. To forward condolences go to www.mcinnisandholloway.com

In loving memory of Makenna Schwebius a tree will be planted at Fish Creek Provincial Park by McInnis &

ENGAGEMENTS/WEDDINGS

Rick & Cheryl Gibeau
and
Werner & Karen Gadosch
proudly and happily
announce the
engagement of their kids,
Farrah Gibeau
and
Mason Gadosch
Bringing our families together.

FIREWORKS
available at
KOOPS SERVICE
Hwy 22, Black Diamond • 933 7242

ANNOUNCEMENTS

Thanks,
From Cubs
2nd & 3rd Scout Troops

To Ray Dymond and Dymond & Safety for donating your time and expertise, in providing the training for our St. John's Ambulance Course.

ANNOUNCEMENTS

Thank You

A heartfelt thanks to family, friends and neighbours for their prayers, flowers, cards, calls, visits and food during and following

ANNOUNCEMENTS

The late JANET MACKAY

We would like to thank the Doctors and Nurses at Oilfields Hospital for the care given to our dear Aunt during her stay, and passing. A special thanks to Leanne Stanley of Snodgrass Funeral Home, Elizabeth Pitura for the lovely service, Connie Hunt for providing the music, our daughter Janice, for the reading, and to all those who attended, sent cards, flowers and to those who shared special memories of Janet.

God Bless you all, Jim & Ethel Wallace

TENDERS

MD OF FOOTHILLS NO. 31
AUDITORS REQUIRED

We are currently accepting quotes from qualified auditors to complete the audit of the Foothills Regional Services Commission. We will require an audit of the consolidated statements for the years ending 2004, 2005 & 2006.

Please submit your tender no later than **4:30 pm on September 22, 2004.**

For further information please contact Christine Hummel at 652-2341

Please submit your tender to

The Foothills Regional Services Commission
Attn: Christine Hummel, CMA
Box 5605
High River, AB T1V 1M7

Figure 6.1.6a *CFT URLs — comparison*

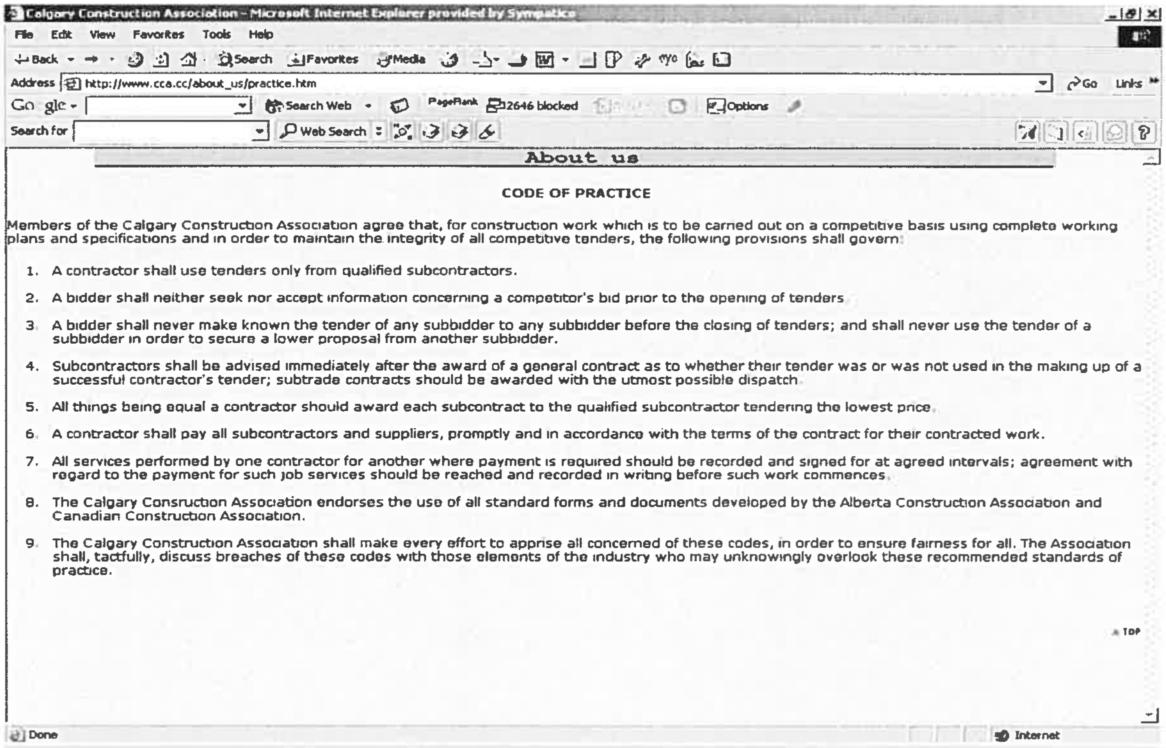
c) same as 4), is a CFT in Calgary;

d) File not found;

e) File not found;

f) See the figure 6.1.6b in next page, it is a "CODE OF PRACTICE" for members of the Calgary Construction Association.

g) same as 6), is a CFT page in Calgary;



Calgary Construction Association - Microsoft Internet Explorer provided by Sympatico

Address http://www.cca.cc/about_us/practice.htm

About us

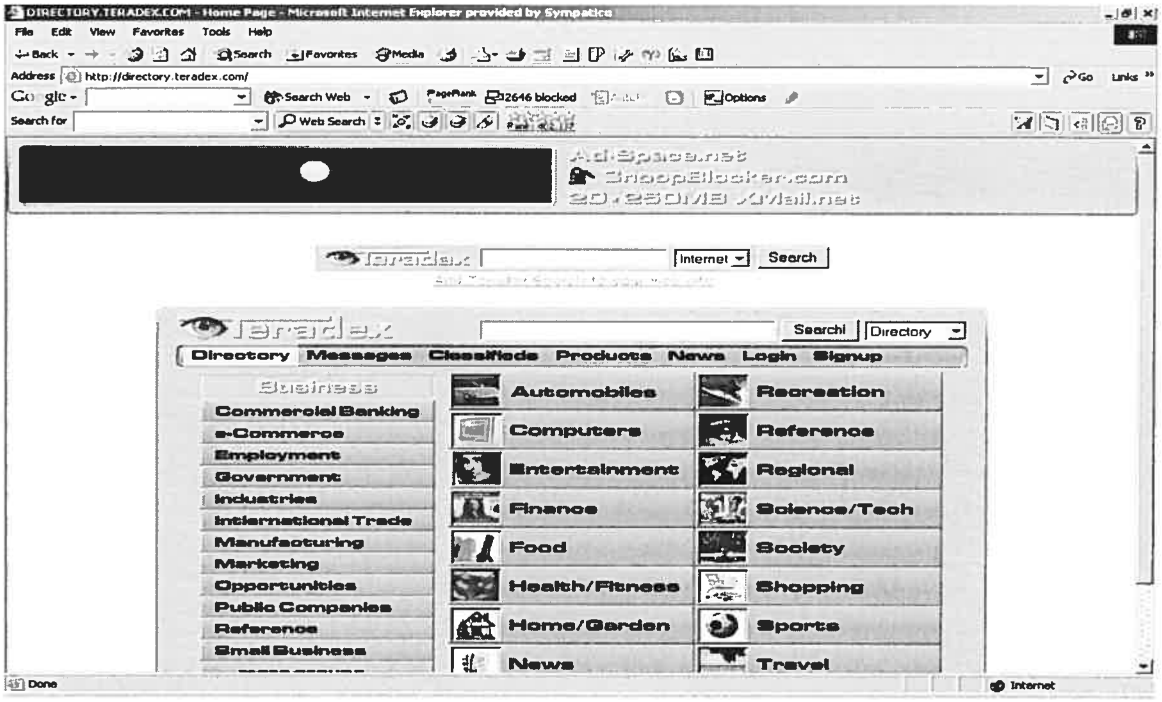
CODE OF PRACTICE

Members of the Calgary Construction Association agree that, for construction work which is to be carried out on a competitive basis using complete working plans and specifications and in order to maintain the integrity of all competitive tenders, the following provisions shall govern:

1. A contractor shall use tenders only from qualified subcontractors.
2. A bidder shall neither seek nor accept information concerning a competitor's bid prior to the opening of tenders.
3. A bidder shall never make known the tender of any subbidder to any subbidder before the closing of tenders; and shall never use the tender of a subbidder in order to secure a lower proposal from another subbidder.
4. Subcontractors shall be advised immediately after the award of a general contract as to whether their tender was or was not used in the making up of a successful contractor's tender; subtrade contracts should be awarded with the utmost possible dispatch.
5. All things being equal a contractor should award each subcontract to the qualified subcontractor tendering the lowest price.
6. A contractor shall pay all subcontractors and suppliers, promptly and in accordance with the terms of the contract for their contracted work.
7. All services performed by one contractor for another where payment is required should be recorded and signed for at agreed intervals; agreement with regard to the payment for such job services should be reached and recorded in writing before such work commences.
8. The Calgary Consruction Association endorses the use of all standard forms and documents developed by the Alberta Construction Association and Canadian Construction Association.
9. The Calgary Construction Association shall make every effort to apprise all concerned of these codes, in order to ensure fairness for all. The Association shall, tactfully, discuss breaches of these codes with those elements of the industry who may unknowingly overlook these recommended standards of practice.

Figure 6.1.6b *CFT URLs — comparison*

h) http://directory.teradex.com/regional/calgary/government is not a CFT page.

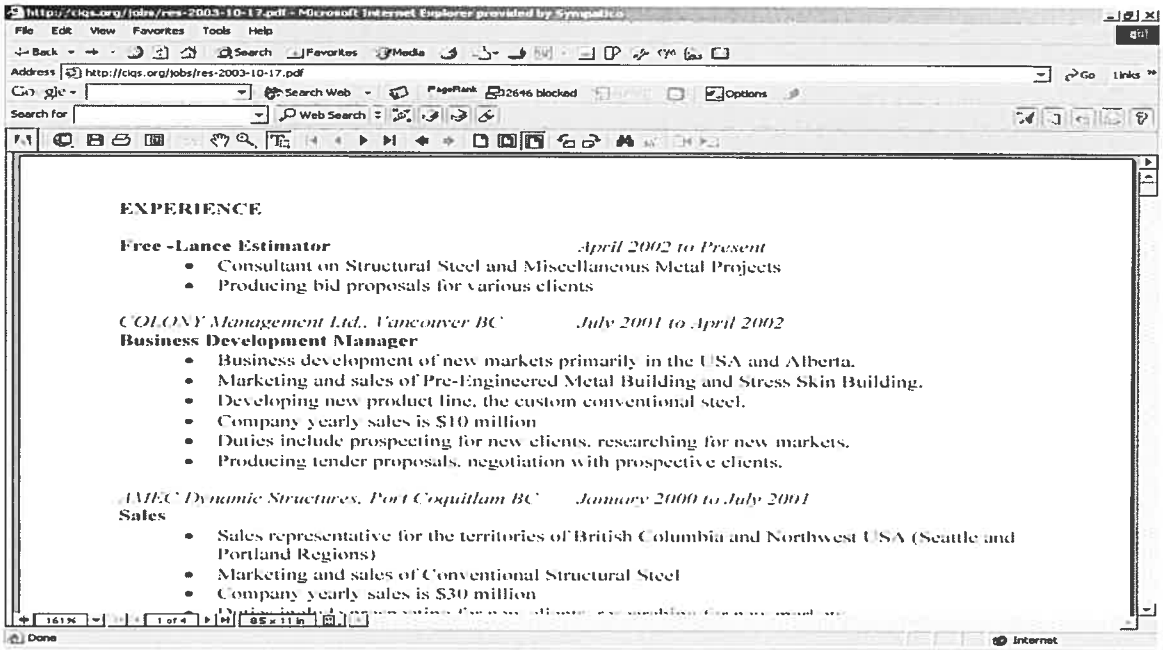


DIRECTORY.TERADEX.COM - Home Page - Microsoft Internet Explorer provided by Sympatico

Address http://directory.teradex.com/

Figure 6.1.6c *CFT URLs — comparison*

i) ciqs.org/jobs/res-2003-10-17.pdf is not a CFT PDF file;


http://ciqs.org/jobs/res-2003-10-17.pdf - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://ciqs.org/jobs/res-2003-10-17.pdf
Go Links
Search Web PageRank 2646 blocked Options
Search for Web Search
161% 1 of 4 8.5 x 11 in
Done
Internet

EXPERIENCE

Free -Lance Estimator *April 2002 to Present*

- Consultant on Structural Steel and Miscellaneous Metal Projects
- Producing bid proposals for various clients

COLONY Management Ltd., Vancouver BC *July 2001 to April 2002*
Business Development Manager

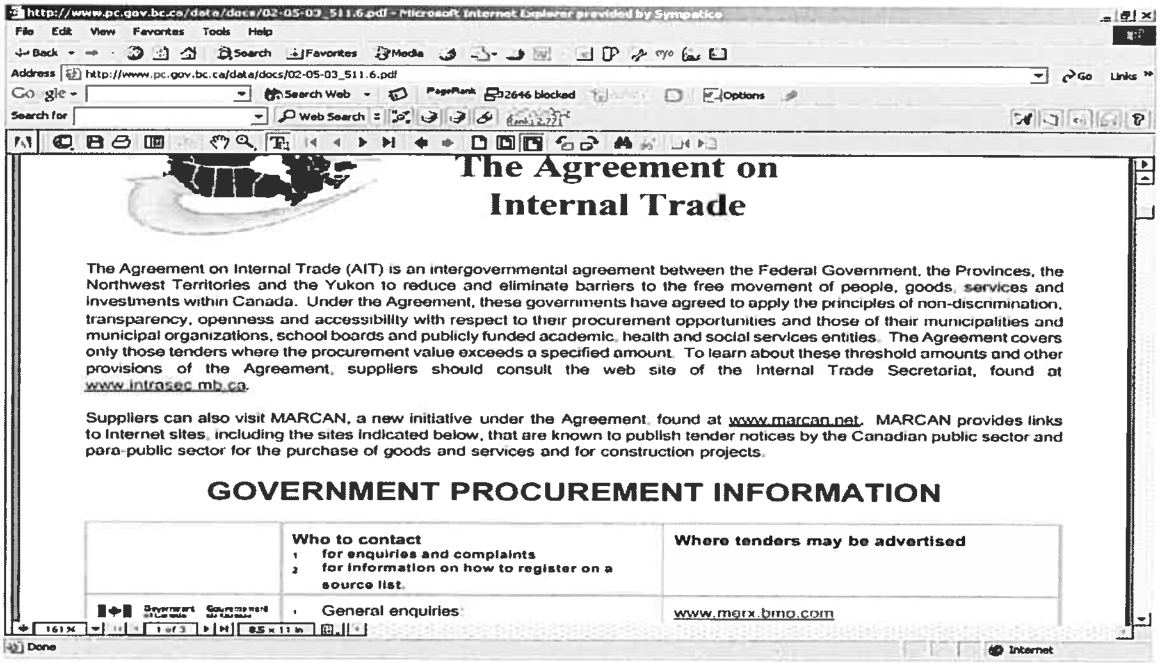
- Business development of new markets primarily in the USA and Alberta.
- Marketing and sales of Pre-Engineered Metal Building and Stress Skin Building.
- Developing new product line, the custom conventional steel.
- Company yearly sales is $10 million
- Duties include prospecting for new clients, researching for new markets.
- Producing tender proposals, negotiation with prospective clients.

AMEC Dynamic Structures, Port Coquitlam BC *January 2000 to July 2001*
Sales

- Sales representative for the territories of British Columbia and Northwest USA (Seattle and Portland Regions)
- Marketing and sales of Conventional Structural Steel
- Company yearly sales is $30 million

Figure 6.1.6d *CFT URLs — comparison*

j) www.pc.gov.bc.ca/data/docs/02-05-03_511.6.pdf is a CFT useful page, but not special for Calgary CFT;


http://www.pc.gov.bc.ca/data/docs/02-05-03_511.6.pdf - Microsoft Internet Explorer provided by Sympatico
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.pc.gov.bc.ca/data/docs/02-05-03_511.6.pdf
Go Links
Search Web PageRank 2646 blocked Options
Search for Web Search
161% 1 of 3 8.5 x 11 in
Done
Internet

The Agreement on Internal Trade

The Agreement on Internal Trade (AIT) is an intergovernmental agreement between the Federal Government, the Provinces, the Northwest Territories and the Yukon to reduce and eliminate barriers to the free movement of people, goods, services and investments within Canada. Under the Agreement, these governments have agreed to apply the principles of non-discnmination, transparency, openness and accessibility with respect to their procurement opportunities and those of their municipalities and municipal organizations, school boards and publicly funded academic, health and social services entities. The Agreement covers only those tenders where the procurement value exceeds a specified amount. To learn about these threshold amounts and other provisions of the Agreement, suppliers should consult the web site of the Internal Trade Secretariat, found at www.intrasec.mb.ca.

Suppliers can also visit MARCAN, a new initiative under the Agreement, found at www.marcan.net. MARCAN provides links to Internet sites, including the sites indicated below, that are known to publish tender notices by the Canadian public sector and para-public sector for the purchase of goods and services and for construction projects.

GOVERNMENT PROCUREMENT INFORMATION

	Who to contact 1 for enquiries and complaints 2 for information on how to register on a source list.	Where tenders may be advertised
Government of Canada Gouvernement du Canada	1 General enquiries:	www.merx.bmo.com

Figure 6.1.6e *CFT URLs — comparison*

Summarizing above top 10 searching results get following table 6.1.2.

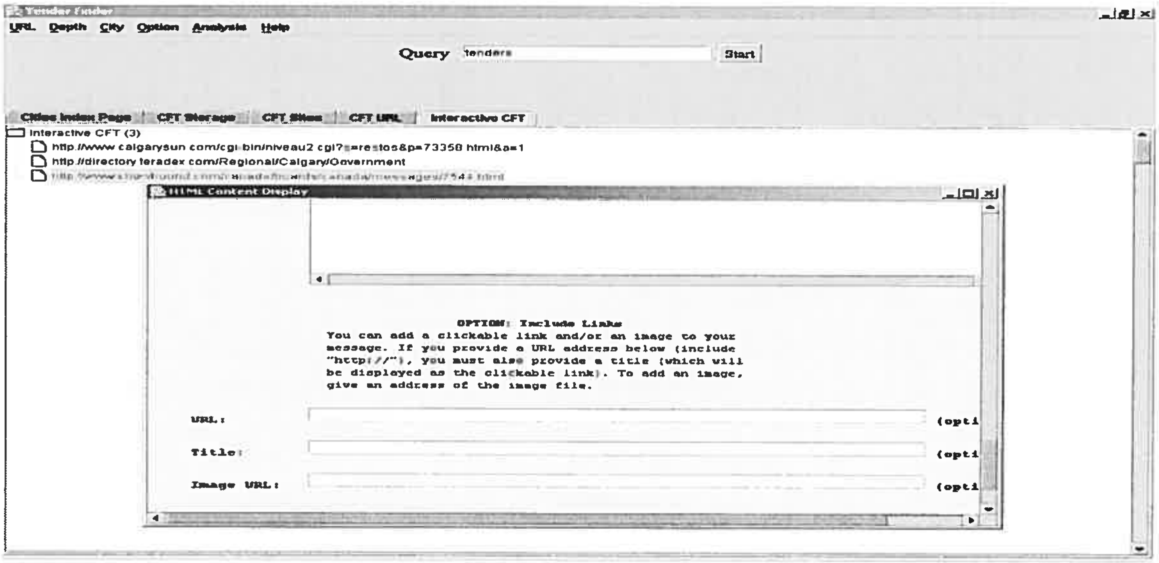
Category	Quantity	Percentage	Comment
CFT in Calgary	3	30%	a),c),g)
Similar to CFT	2	20%	b) in Alberta, j) in Canada
File not Found	2	20%	d),e)
Others	3	30%	f),h),i) less relation to CFT

Table 6.1.2 *Summarizing of evaluation*

6.2.3 Interactive CFT

There are a lot of interactive web pages in CFT domain. Usually they appear as various formats with different forms. It is not easy to auto-fill those forms with an unique way. Anyway, collecting and recording them may help use to grasp some hints for digging further CFT resources.

Interactive CFT same as the CFT URL is a one session searching. At the beginning, a user inputs a query term and presses the "Start" for starting a searching session, after that double click the tabbed menu "Interactive CFT" and wait for a moment. The appearance looks like the following figure.



Query tenders

Interactive CFT (3)

- http://www calgarysun com/cgi-bin/niveau2 cgi?s=restos&p=73358 html&a=1
- http://directory teradex com/Regional/Calgary/Government
- [illegible]

OPTION: Include Links

You can add a clickable link and/or an image to your message. If you provide a URL address below (include "http://"), you must also provide a title (which will be displayed as the clickable link). To add an image, give an address of the image file.

URL: (opti

Title: (opti

Image URL: (opti

Figure 6.1.7 *Interactive CFT example*

Chapter 6 introduced and tested the domain specific system's configurations functions and performances. In chapter 7, I'll give a brief conclusion and some future works to do.

Chapter 7

Conclusions

Performing a search in a specific domain with a traditional, general-purpose search engine is tedious. For this reason, domain specific search engines are becoming increasingly popular, but techniques for successful operation are not fully established. In this thesis, a domain specific searching system (DSS) was developed to complement CSEs (Commercial Search Engine), such as Google and Yahoo. The system employs some specific technologies in web page logical structure analysis and lexical analysis. The prototype has been successful in discovering call for tender pages on the web sites of the cities we tested it with. To achieve this goal, the crawler has to combine many techniques and heuristics. Here are some of the ideas that were found necessary to achieve success:

1) Use of existing search engines to find good starting points for further exploration.
2) Analysis of page structure before examining content. It is critical to distinguish HUB, CONTENT and MIXED pages and to treat each differently.
3) Pages with a complex HTML structure (i.e. Frames and Tables) must be analyzed specially to isolate the information core from repetitive decoration.
4) A keyword profile is important to quickly evaluate the relevance of pages.
5) Development of advanced web software can benefit from the availability of high quality public domain packages.

Preliminary comparisons with Google using additional terms for maximum precision, show that our system's precision is twice as good as Google (about 60% to 30%, see the test data in Chapter 6: Table 6.1.1 & Table 6.1.2). Full evaluation of the performance will have to wait for the next phase of the project; but already it is apparent that the main problem will not be in smart navigation of the web; but rather in establishing classifiers that can detect satisfactorily what is a relevant page and what is not.

Typically all pages found and accepted by our system deal with CFTs in the given city. But most are not CFTs: some are news about contracts, others refer to the terms and conditions of doing business, others are HUBs etc… An automatic classifier is at the heart of any intelligent discovery agent, but more research is required to design a classifier that could learn the subtle differences between these documents.

In retrospect, one search system might function best as a semi-automatic work-bench to allow domain experts to discover sites and write collector scripts for domain specific intelligences gathering systems.

Now that we have a working prototype, we will concentrate on the design of a good classifier as well as implementing the learning techniques proposed by Aggarwal [Aggarwal et al. 2001] in his description of intelligent crawling. We also want to study the best ways to tunnel through to hidden web pages by automatically filling out forms. Finally, we will design experiments to properly evaluate the performance of the system.

Bibliography

[Aas et al. 1999] "Text Categorisation:A Survey" Kjersti Aas and Line Eikvil, http://citeseer.ist.psu.edu/aas99text.html

[Aggarwal et al. 2001] "Intelligent Crawling on the World Wide Web with Arbitrary Predicates", Charu C. Aggarwal, Fatima Al-Garawi & Philip S. Yu, 2001. WWW10 May 1-5,2001. ACM 1-58113-348-0/01/0005

[Apache Jakarta Project 2003] Apache Jakarta Project http://jakarta.apache.org/lucene/docs/index.html

[Ask Jeeves, Inc. 2004] Ask Jeeves, Inc. http://www.ask.com/

[Baroni et al. 2002] "Using textual association measures and minimum edit distance to discover morphological relations", Marco Baroni, Johannes Matiasek & Harald Trost, Collocations Workshop, July,23,2002.

[Bergmark et al. 2002] "Focused Crawls, Tunneling, and Digital Libraries" Donna Bergmark, Carl Lagoze, Alex Sbityakov, in CiteSeer.IST 2002. http://citeseer.ist.psu.edu/bergmark02focused.html

[Bra 1994] P. De Bra, R. Post. Searching for Arbitrary information in the WWW :the Fish-Search for Mosaic. WWW Conference, 1994

[Brin et al. 1998] "The Anatomy of a Large-Scale Hypertextual Web Search Engine" by Sergey Brin and Lawrence Page, 1998

[Carchiolo V. et al. 2003] “Extracting Logical Schema from the Web”, Longheu A. & Malgeri M., in “Applied Intelligence” Vol.18, pp341-355, 200, May, 2003

[Carchiolo et al. 2003] “Extacting Logical Schema form the Web”, Vinceza Carchiolo, Alessandro Longheu and Michele Malgeri, in “Applied Intelligence” Vol. 18, Number 3, May 2003. ISSN:0924-669X

[Chakrabarti et al. 1999] "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery", S. Chakrabarti, M. van den Berg and B. Dom. In Proceedings of the 8th International WWW Conference, Toronto, Canada, May 1999. http://www8.org/w8-papers/5a-search-query/crawling/index.html

[Cho et al. 1998] "Efficient Crawling Through URL Ordering", J. Cho, H. Garcia-Molina, L. Page. In Proceedings of the 7th International WWW Conference, Brisbane, Australia, April 1998. http://www7.scu.edu.au/programme/fullpapers/1919/com1919.htm

[Cho et al. 2002] “Parallel Crawler”, Junghoo Cho & Hector Garcia-Molin, WWW2002, May7-11,2002. ACM 1-58113-449-5/02/0005

[Cleverdon et al. 1966] "Aslib-Cranfield research project",C. W. Cleverdon and E. M. Keen, Cranfield institute of technology, Cranfield; England, , 1966

[Crimmins 2002] “Search System Overview”, Francis Crimmins, 29 May 2002. http://dev.panopticsearch.com/system-overview.html

[F. Paradis et al. 2004] “ MBOI :Scientifique et Technologique (VSST) ” , Toulouse, France, 2004.

[Karger et al. 1999] “Web Caching with Consistent Hashing”, by David Karger, Alex Sherman, Andy Berkheimer, Bill Bogstad, Rizwan Dhanidina, Ken Iwamoto, Brian Kim, Luke Matkins, Yoav Yerushalmi. Published by Elsevier North-Holland, Inc. New York, NY, USA 1999. ISSN:1389-1286

[Kleinberg 1998] J. Kleinber. Authoritative Sources in a Hyperlinked Environment. SODA, 1998.

[Koster 1999] "The Web Robots Pages", M. Koster. 1999. *http://info.webcrawler.com/mak/projects/robots/robots.html*

[Liddy 2001] "How a Search Engine Works" Elizabeth Liddy, May, 2001. FindArticles.com http://www.findarticles.com

[Lancaster,F.W 1968] " Information Retrieval Systems: Characteristics, Testing and Evaluation" published by Wiley, New York, 1968.

[McCallum et al. 1999] "Building Domain-Specific Search Engines with Machine Learning Techniques", A. McCallum, K. Nigam, J. Rennie, and K. Seymore. In 1999 AAAI Spring Symposium on Intelligent Agents in Cyberspace, Stanford University, USA, March 1999. http://www.ri.cmu.edu/pubs/pub_2716.html

[Mladenic 1999] "Text-learning and related intelligent agents" by Dunja Mladenic, in "Applications of Intelligent Information Retrieval", July-August 1999

[Oyama et al. 2003] "Domain-Specific Web Search with Keyword Spices", Satoshi Oyama, Takashi Kokubo, and Toru Ishida, Fellow, IEEE, in IEEE Transaction on knowledge and data engineering, Jan.2004, page(s):17-27,Vol. 16,Issue:1 ISSN: 1041-4347

[Page et al. 1998] "The PageRank citation ranking:Bringing order to the Web", Stanford Digital Library Technologies Project, 1998.

[Sherman et al. 2003] “The Invisible Web”, Chris Sherman &Gary Price, CyberAge Books 0-910965-51-X/softbound, 2003. http://www.searchwise.net/p/iw-fla2003.pdf

[Sullivan April,2003] “Major Search Engines and Directories” Danny Sullivan in “Search Engine Watch” April, 2003 http://www.searchenginewatch.com/links/article.php/2156221

[Sullivan September, 2003] “Search Engine Sizes”, Danny Sullivan, in Search Engine Watch, Sept.2003. http://searchenginewatch.com/reports/article.php/2156481

[Thorsten 1996] “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization (1996)”, Thorsten Joachims in proceedings of ICML-97, 14th International Conference on Machine Learning.

Appendices

1. CFT Feature Words Experiment Data

1) CFT URLs

http://www.oshawa.ca/cit_hall/tenders.asp

http://www.city.brampton.on.ca/purchasing/rfp_2004-057.tml

http://www.city.brampton.on.ca/purchasing/rfp_2004-040.tml

http://www.city.cambridge.on.ca/cs_corporate/purchasing_tenders_list.php

http://www.city.kitchener.on.ca/tenders/tender.asp

http://www.city.london.on.ca/Purchasing/mainpg.htm

http://www.city.niagarafalls.on.ca/cityhall/qtenders.html

http://www.region.peel.on.ca/finance/purchasing/biddocs/index.htm

http://www.city.greatersudbury.on.ca/pubapps/tenders/

http://www.city.toronto.on.ca/tenders/tenders_to.htm

http://www.region.waterloo.on.ca/web/region.nsf/97dfc347666efede85256e590071a3d4/4a3823e3515da81e85256f11004b50c4!OpenDocument

2) Words and Frequencies

Statistic data of all the words and its frequency in the CFT set:

Frequency = 133,	Word = click;		
Frequency = 85,	Word = list;	Frequency = 83,	Word = tender;
Frequency = 76,	Word = acrobat;	Frequency = 75,	Word = format;
Frequency = 74,	Word = pending;	Frequency = 63,	Word = ad;
Frequency = 62,	Word = services;	Frequency = 40,	Word = contract;
Frequency = 40,	Word = addendum;	Frequency = 40,	Word = request;
Frequency = 38,	Word = september;	Frequency = 35,	Word = contact;
Frequency = 35,	Word = add;	Frequency = 35,	Word = mailing;
Frequency = 34,	Word = august;	Frequency = 34,	Word = date;
Frequency = 34,	Word = proposal;	Frequency = 33,	Word = opening;
Frequency = 31,	Word = document;	Frequency = 30,	Word = purchasing;
Frequency = 26,	Word = city;	Frequency = 24,	Word = supply;
Frequency = 23,	Word = issued;	Frequency = 23,	Word = bid;
Frequency = 22,	Word = sep;	Frequency = 22,	Word = peel;
Frequency = 20,	Word = information	Frequency = 24,	Word = tenders;
Frequency = 20,	Word = closed;	Frequency = 20,	Word = works;
Frequency = 19,	Word = wednesday;	Frequency = 19,	Word = aug;
Frequency = 19,	Word = extended;	Frequency = 17,	Word = rfp;
Frequency = 16,	Word = quotation;	Frequency = 16,	Word = plan;
Frequency = 16,	Word = council;	Frequency = 16,	Word = emergency;
Frequency = 16,	Word = construction;	Frequency = 15,	Word = street;

```
Frequency = 14,     Word = deposit;          Frequency = 14,     Word = form;
Frequency = 14,     Word = friday;           Frequency = 14,     Word = replacement;
Frequency = 13,     Word = awarded;          Frequency = 13,     Word = search;
Frequency = 13,     Word = download;          Frequency = 13,     Word = open;
Frequency = 13,     Word = transportation; Frequency = 13,     Word = closing;
Frequency = 13,     Word = buyer;             Frequency = 13,     Word = home;
Frequency = 12,     Word = sept;             Frequency = 12,     Word = proposals;
Frequency = 12,     Word = division;         Frequency = 12,     Word = road;
Frequency = 12,     Word = bids;             Frequency = 12,     Word = site;
Frequency = 12,     Word = brampton;         Frequency = 12,     Word = hall;
Frequency = 12,     Word = perform;          Frequency = 12,     Word = takers;
Frequency = 12,     Word = links;            Frequency = 11,     Word = view;
Frequency = 11,     Word = public;           Frequency = 11,     Word = pdf;
Frequency = 11,     Word = opportunities; Frequency = 14,     Word = regional;
Frequency = 11,     Word = removal;           Frequency = 11,     Word = community;
Frequency = 11,     Word = register;         Frequency = 10,     Word = west;
Frequency = 10,     Word = centre;           Frequency = 10,     Word = delivery;
Frequency = 10,     Word = full;             Frequency = 10,     Word = floor;
Frequency = 10,     Word = fill;             Frequency = 9,     Word = officially;
Frequency = 9,     Word = district;          Frequency = 9,     Word = charge;
Frequency = 9,     Word = mail;              Frequency = 9,     Word = fax;
Frequency = 9,     Word = department;     Frequency = 9,     Word = pwd;
Frequency = 9,     Word = office;            Frequency = 9,     Word = ca;
Frequency = 9,     Word = opportunity;       Frequency = 9,     Word = mississauga;
Frequency = 9,     Word = ontario;           Frequency = 9,     Word = kitchener;
Frequency = 9,     Word = jul;               Frequency = 9,     Word = business;
Frequency = 9,     Word = number;            Frequency = 9,     Word = snow;
Frequency = 8,     Word = quotations;        Frequency = 8,     Word = service;
Frequency = 8,     Word = vendor;            Frequency = 8,     Word = south;
Frequency = 8,     Word = police;            Frequency = 8,     Word = installation;
Frequency = 8,     Word = documents;         Frequency = 7,     Word = sudbury;
Frequency = 7,     Word = october;           Frequency = 7,     Word = summary;
Frequency = 7,     Word = maps;              Frequency = 7,     Word = required;
Frequency = 7,     Word = corporate;          Frequency = 7,     Word = page;
Frequency = 7,     Word = process;           Frequency = 7,     Word = news;
Frequency = 7,     Word = upgrades;           Frequency = 7,     Word = status;
Frequency = 7,     Word = related;           Frequency = 7,     Word = shown;
Frequency = 7,     Word = awards;            Frequency = 7,     Word = health;
Frequency = 6,     Word = bidders;           Frequency = 6,     Word = wellington;
Frequency = 6,     Word = development;    Frequency = 6,     Word = management;
Frequency = 6,     Word = july;              Frequency = 6,     Word = approval;
Frequency = 6,     Word = logo;              Frequency = 6,     Word = web;
Frequency = 6,     Word = provide;           Frequency = 6,     Word = sewer;
Frequency = 6,     Word = housing;           Frequency = 6,     Word = received;
Frequency = 6,     Word = trucks;            Frequency = 6,     Word = sd;
Frequency = 6,     Word = details;           Frequency = 6,     Word = water;
Frequency = 6,     Word = privacy;           Frequency = 6,     Word = oct;
Frequency = 6,     Word = axle;              Frequency = 6,     Word = online;
Frequency = 6,     Word = map;               Frequency = 6,     Word = avenue;
Frequency = 6,     Word = fire;              Frequency = 6,     Word = program;
Frequency = 6,     Word = website;           Frequency = 6,     Word = st;
Frequency = 5,     Word = posted;            Frequency = 5,     Word = accordance;
Frequency = 5,     Word = living;            Frequency = 5,     Word = watermain;
Frequency = 5,     Word = employment;     Frequency = 5,     Word = releases;
Frequency = 5,     Word = cheque;            Frequency = 5,     Word = properties;
Frequency = 5,     Word = calendar;          Frequency = 5,     Word = asphalt;
Frequency = 5,    Word = canada;           Frequency = 5,    Word = questions;
Frequency = 8,    Word = project;          Frequency = 5,    Word = treatment;
Frequency = 5,    Word = email;            Frequency = 5,    Word = economic;
Frequency = 5,    Word = review;           Frequency = 5,    Word = drive;
Frequency = 4,    Word = watermains;       Frequency = 4,    Word = frequently;
Frequency = 4,    Word = time;             Frequency = 4,    Word = region;
Frequency = 4,    Word = jan;              Frequency = 4,    Word = library;
Frequency = 4,    Word = caledon;          Frequency = 4,    Word = results;
Frequency = 4,    Word = sign;             Frequency = 4,    Word = ted;
Frequency = 4,    Word = statement;        Frequency = 4,    Word = rental;
Frequency = 4,    Word = sale;             Frequency = 4,    Word = facilities;
Frequency = 4,    Word = refundable;       Frequency = 4,    Word = nov;
Frequency = 4,    Word = june;             Frequency = 4,    Word = pre;
Frequency = 4,    Word = property;         Frequency = 4,    Word = mar;
Frequency = 4,    Word = jun;              Frequency = 4,    Word = al;
Frequency = 4,    Word = tri;              Frequency = 4,    Word = apr;
Frequency = 4,    Word = improvements; Frequency = 4,    Word = materials;
Frequency = 4,    Word = qualification;    Frequency = 4,    Word = winter;
Frequency = 4,    Word = plant;            Frequency = 4,    Word = directory;
Frequency = 4,    Word = events;           Frequency = 4,    Word = requests;
Frequency = 4,    Word = dec;              Frequency = 4,    Word = thursday;
Frequency = 4,    Word = municipal;        Frequency = 4,    Word = address;
Frequency = 4,    Word = justin;           Frequency = 4,    Word = departments;
Frequency = 4,    Word = rating;           Frequency = 4,    Word = complete;
Frequency = 4,    Word = asked;            Frequency = 4,    Word = hours;
Frequency = 4,    Word = building;         Frequency = 4,    Word = fuel;
Frequency = 4,    Word = maintenance;      Frequency = 4,    Word = wastewater;
Frequency = 4,    Word = ministry;         Frequency = 4,    Word = tickets;
Frequency = 4,    Word = system;           Frequency = 4,    Word = feb;
Frequency = 4,    Word = greater;          Frequency = 4,    Word = current;
Frequency = 4,    Word = rejected;         Frequency = 4,    Word = traffic;
Frequency = 4,    Word = official;         Frequency = 4,    Word = general;
Frequency = 4,    Word = livingston;       Frequency = 4,    Word = programs;
Frequency = 4,    Word = procedure;        Frequency = 4,    Word = storm;
Frequency = 3,    Word = phone;            Frequency = 3,    Word = fee;
Frequency = 3,    Word = cash;             Frequency = 3,    Word = listed;
Frequency = 3,    Word = call;             Frequency = 3,    Word = install;
Frequency = 3,    Word = mayor;            Frequency = 3,    Word = projects;
Frequency = 3,    Word = file;             Frequency = 3,    Word = hot;
Frequency = 3,    Word = copy;             Frequency = 3,    Word = include;
Frequency = 3,    Word = refer;            Frequency = 3,    Word = interest;
Frequency = 3,    Word = tr;               Frequency = 3,    Word = cold;
Frequency = 3,    Word = law;              Frequency = 3,    Word = square;
Frequency = 3,    Word = main;             Frequency = 3,    Word = computer;
Frequency = 3,    Word = taxation;         Frequency = 3,    Word = bridge;
Frequency = 3,    Word = accepted;         Frequency = 3,    Word = working;
Frequency = 3,    Word = grants;           Frequency = 3,    Word = souza;
Frequency = 3,    Word = corporation;      Frequency = 3,    Word = facility;
Frequency = 3,    Word = registration;     Frequency = 3,    Word = control;
Frequency = 3,    Word = order;            Frequency = 3,    Word = live;
Frequency = 3,    Word = operative;        Frequency = 3,    Word = long;
Frequency = 3,    Word = repairs;          Frequency = 3,    Word = social;
Frequency = 3,    Word = march;            Frequency = 3,    Word = submitted;
Frequency = 3,    Word = exterior;         Frequency = 3,    Word = heritage;
Frequency = 3,    Word = close;            Frequency = 3,    Word = power;
Frequency = 3,    Word = financial;        Frequency = 3,    Word = column;
Frequency = 3,    Word = telephone;       Frequency = 3,    Word = affordable;
Frequency = 3,    Word = yard;            Frequency = 3,    Word = submit;
Frequency = 3,    Word = local;           Frequency = 3,    Word = items;
Frequency = 3,    Word = vendors;         Frequency = 3,    Word = esd;
Frequency = 3,    Word = renovations;     Frequency = 3,    Word = intersection;
Frequency = 3,    Word = purchase;        Frequency = 3,    Word = arena;
Frequency = 3,    Word = parking;         Frequency = 3,    Word = supplies;
Frequency = 3,    Word = jobs;            Frequency = 3,    Word = note;
Frequency = 3,    Word = publications;    Frequency = 3,    Word = planning;
Frequency = 3,    Word = work;            Frequency = 3,    Word = fleet;
Frequency = 3,    Word = rehabilitation;  Frequency = 3,    Word = signed;
Frequency = 3,    Word = care;            Frequency = 3,    Word = citizen;
Frequency = 3,    Word = recreation;      Frequency = 3,    Word = advanced;
Frequency = 3,    Word = recycling;       Frequency = 3,    Word = quotes;
Frequency = 3,    Word = archives;        Frequency = 3,    Word = disclaimer;
Frequency = 3,    Word = systems;         Frequency = 3,    Word = certificate;
Frequency = 3,    Word = provided;        Frequency = 3,    Word = fair;
Frequency = 2,    Word = digital;         Frequency = 2,    Word = top;
Frequency = 2,    Word = visual;          Frequency = 2,    Word = expression;
Frequency = 2,    Word = operation;       Frequency = 2,    Word = payable;
Frequency = 2,    Word = section;         Frequency = 2,    Word = depot;
Frequency = 2,    Word = trunk;           Frequency = 2,    Word = provision;
Frequency = 2,    Word = times;           Frequency = 2,    Word = child;
Frequency = 2,    Word = diesel;          Frequency = 2,    Word = supplier;
Frequency = 2,    Word = owned;           Frequency = 2,    Word = manor;
Frequency = 2,    Word = decisions;       Frequency = 2,    Word = thedomain;
Frequency = 2,    Word = biodiesel;       Frequency = 2,    Word = prequalification;
Frequency = 2,    Word = staff;           Frequency = 2,    Word = finance;
Frequency = 2,    Word = sports;          Frequency = 2,    Word = visitors;
Frequency = 2,    Word = important;       Frequency = 2,    Word = schedule;
Frequency = 2,    Word = receiving;       Frequency = 2,    Word = hamilton;
Frequency = 2,    Word = onmousedown; Frequency = 2,    Word = tom;
Frequency = 2,    Word = nile;            Frequency = 2,    Word = animal;
Frequency = 2,    Word = groups;          Frequency = 2,    Word = data;
Frequency = 2,    Word = wards;           Frequency = 2,    Word = sidewalks;
Frequency = 2,    Word = guide;           Frequency = 2,    Word = planting;
Frequency = 2,    Word = reserved;        Frequency = 2,    Word = roadwatch;
Frequency = 2,    Word = streets;         Frequency = 2,    Word = plowing;
Frequency = 2,    Word = qualifications;  Frequency = 2,    Word = contents;
Frequency = 2,    Word = responsibility;  Frequency = 2,    Word = interior;
Frequency = 2,    Word = calling;         Frequency = 2,    Word = east;
Frequency = 2,    Word = accommodations; Frequency = 2,  Word = rights;
Frequency = 2,    Word = restoration;     Frequency = 2,    Word = shepherd;
Frequency = 2,    Word = gm;              Frequency = 2,    Word = solicitation;
Frequency = 2,    Word = technology;      Frequency = 2,    Word = agent;
Frequency = 2,    Word = français;        Frequency = 2,    Word = mapping;
Frequency = 2,    Word = roads;           Frequency = 2,    Word = successes;
Frequency = 2,    Word = event;           Frequency = 2,    Word = bidding;
Frequency = 2,    Word = wage;            Frequency = 2,    Word = ed;
Frequency = 2,    Word = originally;      Frequency = 2,    Word = prices;
Frequency = 2,    Word = title;           Frequency = 2,    Word = transhelp;
Frequency = 2,    Word = specific;        Frequency = 2,    Word = check;
Frequency = 2,    Word = mix;             Frequency = 2,    Word = certified;
Frequency = 2,    Word = dump;            Frequency = 2,    Word = marked;
Frequency = 2,    Word = measures;        Frequency = 2,    Word = human;
Frequency = 2,    Word = tt;              Frequency = 2,    Word = insurance;
Frequency = 2,   Word = oxford;           Frequency = 2,   Word = newspaper;
Frequency = 2,   Word = capreol;          Frequency = 2,   Word = formal;
Frequency = 2,   Word = rec;              Frequency = 2,   Word = arts;
Frequency = 2,   Word = moonah;           Frequency = 2,   Word = minutes;
Frequency = 2,   Word = lakeview;         Frequency = 2,   Word = transit;
Frequency = 2,   Word = tax;              Frequency = 2,   Word = require;
Frequency = 2,   Word = reference;        Frequency = 2,   Word = king;
Frequency = 2,   Word = celebrate;        Frequency = 2,   Word = storage;
Frequency = 2,   Word = architectural;    Frequency = 2,   Word = entrance;
Frequency = 2,   Word = submission;       Frequency = 2,   Word = international;
Frequency = 2,   Word = dundas;           Frequency = 2,   Word = hardware;
Frequency = 2,   Word = davies;           Frequency = 2,   Word = term;
Frequency = 2,   Word = consulting;       Frequency = 2,   Word = faq;
Frequency = 2,   Word = clarkson;         Frequency = 2,   Word = line;
Frequency = 2,   Word = lowest;           Frequency = 2,   Word = copyright;
Frequency = 2,   Word = application;      Frequency = 2,   Word = irene;
Frequency = 2,   Word = downtown;         Frequency = 2,   Word = envelope;
Frequency = 2,   Word = tandem;           Frequency = 2,   Word = attractions;
Frequency = 2,   Word = pioneer;          Frequency = 2,   Word = notices;
Frequency = 2,   Word = instructions;     Frequency = 2,   Word = make;
Frequency = 2,   Word = pick;             Frequency = 2,   Word = initiatives;
Frequency = 2,   Word = toronto;          Frequency = 2,   Word = landscape;
Frequency = 2,   Word = murphy;           Frequency = 2,   Word = joint;
Frequency = 2,   Word = performing;       Frequency = 2,   Word = hss;
Frequency = 2,   Word = laws;             Frequency = 2,   Word = hill;
Frequency = 2,   Word = sealed;           Frequency = 2,   Word = statements;
Frequency = 2,   Word = software;         Frequency = 2,   Word = advertised;
Frequency = 2,   Word = virus;            Frequency = 2,   Word = monday;
Frequency = 2,   Word = alterations;      Frequency = 2,   Word = windows;
Frequency = 2,   Word = high;             Frequency = 2,   Word = bidder;
Frequency = 2,   Word = policy;           Frequency = 2,   Word = place;
Frequency = 2,   Word = media;            Frequency = 2,   Word = wp;
Frequency = 2,   Word = cls;              Frequency = 2,   Word = theatre;
Frequency = 2,   Word = shane;            Frequency = 2,   Word = electrical;
Frequency = 2,   Word = training;         Frequency = 2,   Word = asphaltic;
Frequency = 2,   Word = competitive;      Frequency = 2,   Word = agendas;
Frequency = 2,   Word = area;             Frequency = 2,   Word = theemail;
Frequency = 2,   Word = mechanical;       Frequency = 2,   Word = bloom;
Frequency = 2,   Word = box;              Frequency = 2,   Word = communities;
Frequency = 2,   Word = funds;            Frequency = 2,   Word = resources;
Frequency = 2,   Word = cao;              Frequency = 2,   Word = checkbutton;
Frequency = 2,   Word = landfill;         Frequency = 2,   Word = phase;
Frequency = 2,   Word = limited;          Frequency = 2,   Word = tourism;
Frequency = 2,   Word = technical;        Frequency = 2,   Word = mentioned;
Frequency = 2,   Word = images;           Frequency = 2,   Word = reports;
Frequency = 2,   Word = surplus;          Frequency = 2,   Word = hospital;
Frequency = 2,   Word = opened;           Frequency = 2,   Word = peelregion;
Frequency = 2,   Word = land;             Frequency = 2,   Word = design;
Frequency = 2,   Word = pollution;        Frequency = 2,   Word = mailed;
Frequency = 2,   Word = dearness;         Frequency = 2,   Word = courier;
Frequency = 2,   Word = company;          Frequency = 2,   Word = december;
Frequency = 2,   Word = porcarelli;       Frequency = 2,   Word = entry;
Frequency = 2,   Word = reconstruction;   Frequency = 2,   Word = leisure;
Frequency = 2,   Word = aluminum;         Frequency = 2,   Word = signage;
Frequency = 2,   Word = seasonal;         Frequency = 2,   Word = procedures;
Frequency = 2,   Word = canadian;         Frequency = 2,   Word = parks;
Frequency = 2,    Word = gas;               Frequency = 2,    Word = calls;
Frequency = 2,    Word = gr;                Frequency = 2,    Word = master;
Frequency = 2,    Word = reader;            Frequency = 2,    Word = necessarily;
Frequency = 2,    Word = sand;              Frequency = 2,    Word = rfps;
Frequency = 2,    Word = tourists;          Frequency = 1,    Word = tracing;
Frequency = 1,    Word = wl;                Frequency = 1,    Word = tw;
Frequency = 1,    Word = budgets;           Frequency = 1,    Word = notified;
Frequency = 1,    Word = paid;              Frequency = 1,    Word = taxes;
Frequency = 1,    Word = purchasingpolicy;Frequency = 1,    Word = village;
Frequency = 1,    Word = greatersudbury;   Frequency = 1,    Word = shelter;
Frequency = 1,    Word = energy;            Frequency = 1,    Word = publicly;
Frequency = 1,    Word = tool;              Frequency = 1,    Word = explorer;
Frequency = 1,    Word = location;          Frequency = 1,    Word = body;
Frequency = 1,    Word = fail;              Frequency = 1,    Word = chief;
Frequency = 1,    Word = extension;         Frequency = 1,    Word = administrative;
Frequency = 1,    Word = located;           Frequency = 1,    Word = sunforest;
Frequency = 1,    Word = wheelchair;        Frequency = 1,    Word = easan;
Frequency = 1,    Word = terrace;           Frequency = 1,    Word = lookout;
Frequency = 1,    Word = bank;              Frequency = 1,    Word = pump;
Frequency = 1,    Word = oshawa;            Frequency = 1,    Word = de;
Frequency = 1,    Word = residents;         Frequency = 1,    Word = singles;
Frequency = 1,    Word = railway;           Frequency = 1,    Word = fiber;
Frequency = 1,    Word = accessibility;     Frequency = 1,    Word = urb;
Frequency = 1,    Word = speed;             Frequency = 1,    Word = core;
Frequency = 1,    Word = courts;            Frequency = 1,    Word = norris;
Frequency = 1,    Word = assessment;        Frequency = 1,    Word = fees;
Frequency = 1,    Word = comment;           Frequency = 1,    Word = surina;
Frequency = 1,    Word = pay;               Frequency = 1,    Word = viewed;
Frequency = 1,    Word = incomplete;        Frequency = 1,    Word = server;
Frequency = 1,    Word = introduction;      Frequency = 1,    Word = bovaird;
Frequency = 1,    Word = installed;         Frequency = 1,    Word = 0dz7udyj4;
Frequency = 1,    Word = untitled;          Frequency = 1,    Word = audit;
Frequency = 1,    Word = civic;             Frequency = 1,    Word = sales;
Frequency = 1,    Word = trades;            Frequency = 1,    Word = consultant;
Frequency = 1,    Word = enterprises;       Frequency = 1,    Word = playground;
Frequency = 1,    Word = symbol;            Frequency = 1,    Word = acceptable;
Frequency = 1,    Word = cable;             Frequency = 1,    Word = cellular;
Frequency = 1,    Word = assorted;          Frequency = 1,    Word = finlayson;
Frequency = 1,    Word = middleton;         Frequency = 1,    Word = abc;
Frequency = 1,    Word = steeles;           Frequency = 1,    Word = parts;
Frequency = 1,    Word = curb;              Frequency = 1,    Word = valves;
Frequency = 1,    Word = heat;              Frequency = 1,    Word = caradoc;
Frequency = 1,    Word = environment;      Frequency = 1,    Word = utility;
Frequency = 1,    Word = apparatus;         Frequency = 1,    Word = successful;
Frequency = 1,    Word = north;             Frequency = 1,    Word = internet;
Frequency = 1,    Word = cancelled;         Frequency = 1,    Word = jalna;
Frequency = 1,    Word = vehicles;          Frequency = 1,    Word = industry;
Frequency = 1,    Word = ew;                Frequency = 1,    Word = price;
Frequency = 1,    Word = visa;              Frequency = 1,    Word = logging;
Frequency = 1,    Word = send;              Frequency = 1,    Word = don;
Frequency = 1,    Word = archive;           Frequency = 1,    Word = valley;
Frequency = 1,    Word = airport;           Frequency = 1,    Word = crime;
Frequency = 1,    Word = leo;               Frequency = 1,    Word = rob;
Frequency = 1,    Word = language;          Frequency = 1,    Word = electronic;
Frequency = 1,    Word = ski;               Frequency = 1,    Word = roof;
Frequency = 1,    Word = marriage;          Frequency = 1,    Word = outfall;
```

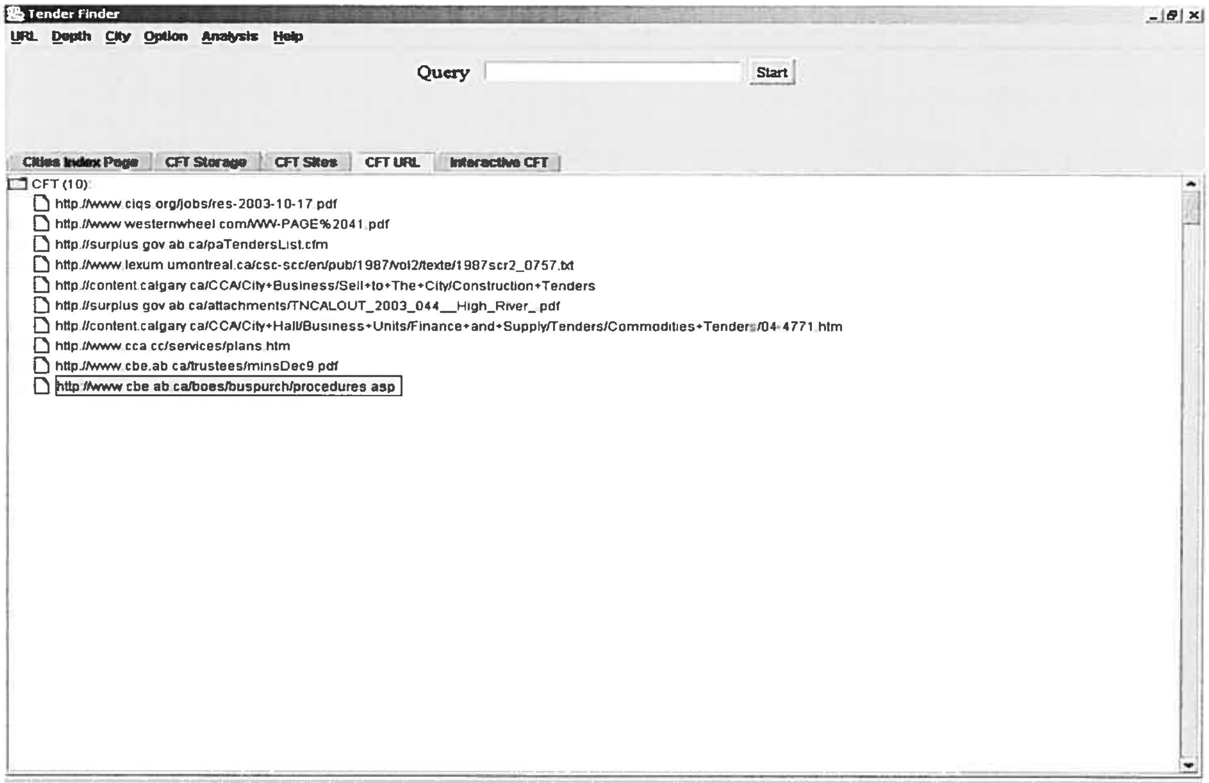
Frequency = 1, Word = therapy; Frequency = 1, Word = walden;
Frequency = 1, Word = workplace;

2. CFT Searching in Five Cities

(Yes: CFT pertinent; No: not a CFT page; Crawler depth:2; Page:2)

Calgary

http://www.ciqs.org/jobs/res-2003-10-17.pdf No
http://www.westernwheel.com/WW-PAGE%2041.pdf Yes
http://surplus.gov.ab.ca/paTendersList.cfm Yes
http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/vol2/texte/1987scr2_0757.txt No
http://content.calgary.ca/CCA/City+Business/Sell+to+The+City/Construction+Tenders Yes
http://surplus.gov.ab.ca/attachments/TNCALOUT_2003_044__High_River_.pdf Yes
http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders/Commodities+Tenders/04-4771.htm Yes
http://www.cca.cc/services/plans.htm Yes
http://www.cbe.ab.ca/trustees/minsDec9.pdf
http://www.cbe.ab.ca/boes/buspurch/procedures.asp Yes



```
CFT (10)
  http://www.ciqs.org/jobs/res-2003-10-17.pdf
  http://www.westernwheel.com/WW-PAGE%2041.pdf
  http://surplus.gov.ab.ca/paTendersList.cfm
  http://www.lexum.umontreal.ca/csc-scc/en/pub/1987/vol2/texte/1987scr2_0757.txt
  http://content.calgary.ca/CCA/City+Business/Sell+to+The+City/Construction+Tenders
  http://surplus.gov.ab.ca/attachments/TNCALOUT_2003_044__High_River_.pdf
  http://content.calgary.ca/CCA/City+Hall/Business+Units/Finance+and+Supply/Tenders/Commodities+Tenders/04-4771.htm
  http://www.cca.cc/services/plans.htm
  http://www.cbe.ab.ca/trustees/minsDec9.pdf
  http://www.cbe.ab.ca/boes/buspurch/procedures.asp
```

Appendix Figure *Test in Calgary*

2) Vancouver

http://www.translink.bc.ca/About_TransLink/Business_Opportunities/Bid_Details/Q4-0014.asp Yes

http://www.crd.bc.ca/tenders/041510_trenchless_pipe_rehab.pdf Yes

http://www.bchydro.com/rx_files/info/info13924.pdf Yes

http://www.gvrd.bc.ca/gvrdtenders/TenderDetails.aspx?tenderId=274&tenderType=open Yes

http://www.gvrd.bc.ca/gvrdtenders/TenderDetails.aspx?tenderId=279&tenderType=open Yes

http://www.grainscanada.gc.ca/newsroom/misc/tenders/surplusvan-e.htm Yes

http://www.dnv.org/upload/documents/cpolicy/c209501.pdf Yes

http://www.bestwesterncoqinn.com/pantrylunch.html No

http://www.bchydro.com/rx_files/info/info9912.pdf No

http://www.city.vancouver.bc.ca/ctyclerk/ADS/ADcontract2004-01.htm Yes

http://www.885westgeorgia.com/extranet/news_emp.html No

http://www.city.richmond.bc.ca/webnews/tenders/docs/0818_park1570.pdf Yes

http://www.translink.bc.ca/About_TransLink/Business_Opportunities/default.asp Yes

3) Winnipeg

http://www.mpi.mb.ca/salvage/auctionprint.asp?salenm=156 Yes

http://www.winnipeg.ca/matmgt Yes

http://www.mpi.mb.ca/salvage/auctionprint.asp?salenm=167 Yes

http://www.winnipeg.ca/finance/findata/matmgt/documents/2004/196-2004/196-2004_Part_A-Submission.pdf Yes

http://www.prha.mb.ca/news.htm CFT News No.

http://www.mhca.mb.ca/HeavyNews/heavy_news_june3_04.pdf No

http://www.winnipeg-chamber.com/pdf/policy/fldsol.pdf Yes

http://www.smartwinnipeg.mb.ca/Smart_Recommendations.htm CFT News, No

http://www.cwb.ca/en/publications/farmers/pdf/nov-dec-2003.pdf No

http://www.gov.mb.ca/tgs/gs/contracts/tenders/ads/1198_04_ad.pdf Yes

http://www.wsd1.org/Board/policy_minutes/Policy/Policy_DJC.pdf Yes http://www.cwb.ca/en/publications/students_researchers/pdf/2002-03_full_english_statistics.pdf No

http://www.cwb.ca/en/movement/elevator_managers/pdf/2004-05_elevatorguide.pdf No

http://www.mhca.mb.ca/HeavyNews/heavy_news_april29.pdf Yes

http://www.canoe.ca/NewsStand/Columnists/Winnipeg/Frank_Landry/2004/07/30/562424.html No

http://www.canadinns.com/confbanq/ambmenuwin.pdf No

http://www.gov.mb.ca/tgs/gs/contracts/tenders/ads/0212sr04_ad.pdf Yes

3) Yellowknife

http://www.canlii.org/nt/laws/regu/f-3/20040512/whole.html Yes

http://www.recycle.ab.ca/2003Proceedings/JaredBuchko.pdf No

http://www.amec.com/earthandenvironmental/where/Officeprofiles.asp?PageID=1081&OfficeID=88 No

http://www.ozemail.com.au/~kshapley/flash.html No

http://city.yellowknife.nt.ca/userfiles/page_attachments/Library/1/358200_LAND_ADMINISTRATION_BY-LAW_NO_3853.pdf Yes

http://www.diavik.ca/dialogue/dialogue15.pdf No

http://collection.nlc-bnc.ca/100/201/300/first_perspective/2001/05-10/employment/emp1050102.html Yes

http://www.yukonweb.com/community/yukon-news/apr5.htmld Yes

http://www.wcb.nt.ca/AboutWCB/pdf/Goverance%20Council/Dec2002.pdf No

http://city.yellowknife.nt.ca/userfiles/page_attachments/Library/1/3143357_CAPITAL_UPDATE_JULY_30__2004_-_version_2.pdf Yes

http://www.pwgsc.gc.ca/reports/text/pdfs/DPR_PWGSC_2003_e.pdf Yes

http://www.nnsl.com/yir/yir01/yirdehcho01.html No

http://www.cbsc.org/english/search/display.cfm?code=4066&coll=FE_FEDSBIS_E Yes

4) Halifax

http://www.hrwc.ns.ca/whats_new/tenders.html Yes

http://www.bids.ca/protect/tndr290.htm Yes

http://www.herald.ns.ca/stories/2004/09/07/fEntertainment127.raw.html No

http://www.region.halifax.ns.ca/procurement/tenders3.asp Yes

http://www.novaforestalliance.com/media/documents/tender_offer_Mar26-02.pdf Yes

http://www.acoa.ca/e/media/press/press.shtml?1401 Yes

http://www.nseia.ns.ca/EnviroNews%20%20March%2026,%202004.pdf No

http://www.cbsc.org/english/search/display.cfm?code=4066&coll=FE_FEDSBIS_E Yes

http://www.bids.ca/protect/tndr670.htm Yes

5) Ajax

http://www.georgeortiz.com/3D/220 NO

http://www.ipc.on.ca/scripts/index_.asp?action=31&P_ID=7589&N_ID=1&PT_ID=923&U_ID=0 Yes

http://www.townofajax.com/government/minutes/2004/gg-july8.pdf No

http://www.trainandemploy.qld.gov.au/client/resources/about/research_publications/strategy_policy/pdf/example_plan.pdf Yes

http://www.bids.ca/tenderlist_agency.php?AGENCY=AJAX Yes

http://news.moneycentral.msn.com/ticker/sigdev.asp?Symbol=IBN CFT report No

http://www.rodney.govt.nz/council/minutes/fullcouncilsuppagenda29JAN04.pdf Yes

http://news.bbc.co.uk/sport/hi/english/football/champions_league/newsid_1449000/1449507.stm No