World Wide Web Site Summarization

Yongzheng Zhang
Nur Zincir-Heywood
Evangelos Milios

Technical Report CS-2002-08

January 10, 2002

Faculty of Computer Science
6050 University Ave., Halifax, Nova Scotia, B3H 1W5, Canada

World Wide Web Site Summarization

Yongzheng Zhang, Nur Zincir-Heywood, Evangelos Milios

Faculty of Computer Science,

Dalhousie University,

Halifax, N.S.,

Canada B3H 1W5

{yongzhen,zincir,eem@cs.dal.ca}

October 10, 2002

Abstract

As the size and diversity of the World Wide Web grows rapidly, it is becoming more and more difficult for a user to skim over a Web site and get an idea of its contents. Currently, manually constructed summaries by volunteer experts are available, such as the DMOZ Open Directory Project. This research is directed towards automating the Web site summarization task. To achieve this objective, an approach which applies machine learning and natural language processing techniques is developed to summarize a Web site automatically. The information content of the automatically generated summaries is compared, via a formal evaluation process involving human subjects, to DMOZ summaries, home page browsing and time-limited site browsing, for a number of academic and commercial Web sites. Statistical evaluation of the scores of the answers to a list of questions about the sites demonstrates that the automatically generated summaries convey the same information to the reader as DMOZ summaries do, and more information than the two browsing options.

1 Introduction

As the size and diversity of the World Wide Web grows rapidly, vast amounts of online information continues to grow at an incredible rate, leading to "information overload" [21]. It has been more and more difficult for the user to skim over a Web site and get an idea of its contents. Currently, manually constructed summaries by volunteer experts are available, such as the DMOZ Open Directory Project [1]. These human-authored summaries give a concise and effective description of popular Web sites. However, they are subjective, and expensive to build and maintain [8]. Our objective is to summarize the Web site automatically.

The information overload problem has brought users great difficulty to find useful information quickly and effectively. The technology of automatic summarization of text is maturing and may provide a solution to this problem [21, 19]. Automatic text summarization produces a concise summary by abstraction or extraction of important text using statistical approaches [9], linguistic approaches [4] or combination of the two [5, 14, 19].

The goal of abstraction is to produce coherent summaries that are as good as human authored summaries. However, this is very difficult to achieve with current natural language processing techniques [14]. During the last few years, extraction techniques have been the focus of automatic text summarization research [17]. Extraction systems analyze a source document to determine significant sentences, and produce a concise summary from these significant sentences. The significance of a sentence is determined by a few features such as the density of keywords and rhetorical relations in the context [26].

With the continuing explosion of online data, demand for multi-document summarization techniques is growing. Multi-document summaries with high quality can save users much

time in reading relevant text documents or browsing interesting Web sites. Many of the single-document summarization techniques can also be used in multi-document summarization. Moreover, features such as information about the whole document set and relationships between individual documents can also be analyzed and incorporated in the multi-document summarization [15, 19].

There are two major approaches to summarization evaluation: *intrinsic* and *extrinsic* [16, 21]. Intrinsic evaluation compares automatically generated summaries with a gold standard (ideal summaries). Extrinsic evaluation measures the performance of automatically created summaries in a particular task (e.g., classification). The extrinsic evaluation is also called task-based evaluation and it has become more and more popular recently [23]. In recent years, the importance o evaluation of summarization algorithms has increased (e.g. [20, 13]), due to the desire for better summaries [19].

Research interest in Web page summarization has been increasing over the last several years. Basically Web page summarization derives from text summarization techniques [9]. However, it is a great challenge to summarize Web pages automatically and effectively [3], because Web pages differ from traditional text documents in both structure and content. Instead of coherent text with a well-defined discourse structure, Web pages often have diverse contents such as bullets and images [6].

Currently there is no effective way to produce unbiased, coherent and informative summaries of Web pages automatically. Amitay et al [3] propose a unique approach, which relies on the hypertext structure and the way information is described under it. Instead of analyzing the document, this approach exploits Web authoring conventions, understanding how useful information can be extracted from hypertext layout and language structure. This

approach is applied to "generate short coherent textual snippets presented to the user with search engine results" [3].

Garcia-Molina et al [9] compare alternative methods to summarize Web pages for display on handheld devices. The *Keyword* method extracts keywords from the text units, and the *Summary* method identifies the most significant sentence of each text unit as a summary for the unit. They test the performance of these methods by asking human subjects to perform specific tasks using each method, and conclude that the combined *Keyword/Summary* method provides the best performance in terms of access times and number of pen actions on the hand held devices.

Our objective is to automate summarization of Web sites, not simply Web pages. To this end, the "Keyword/Summary" idea of [9] is adopted. However, this methodology is enhanced by applying machine learning and natural language processing techniques. A summary is produced in a sequence of stages as indicated in Figure 1. First a given number of web pages is collected from a given web site via a breadth-first search starting at the home page. Second, plain text is extracted from these pages and partitioned into text paragraphs by the text browser *Lynx* (Section 2). In the third step, *short* paragraphs are first filtered out of all paragraphs, then *long* paragraphs are classified into *narrative* or *non-narrative*. Criteria for these two operations were achieved by training data sets using both machine learning approaches and natural language processing techniques (Section 3). Fourth, decision tree rules are used to extract key-phrases from narrative text, anchor text and special text (e.g., italic text), separately. These rules were achieved by applying machine learning approaches to measure how significant each category of key-phrases is (Section 4). After key-phrases are extracted, *narrative* paragraphs will be reviewed in order to find the

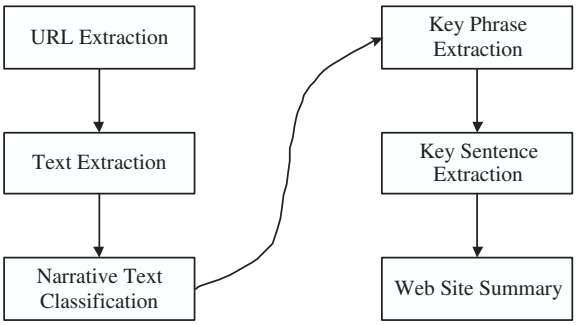

Figure 1: Web site summarization process

most significant sentences, as the ones containing a high density of key-phrases. Finally, the summary is generated, which consists of the top 25 key-words, top 10 key-terms and top 5 key-sentences (Section 5). The automatically generated summaries are compared with the DMOZ summaries of the same site, home page browsing and time-limited site browsing, for a number of academic and commercial Web sites (Section 6).

2 Web Page and Text Extraction

In general, the structure of a Web site is hierarchical. In a breadth-first traversal of a Web site, the home page is the root, i.e., first level. All Web pages pointed at from the home page are in the second level, and so on. Intuitively, contents of pages near the top are more representative of the general content of the site than pages deeper into the site. The home page of a Web site often presents a brief description of what this site is about. When we go deeper into the Web site, Web pages tend to discuss specific topics in detail.

Since our objective is to summarize the Web site, we want to focus on top-level pages in order to extract the contents which describe the Web site in a general sense. A module called

Site Crawler was developed, that crawls within a given web site using *breadth-first-search*. This means that only Web pages physically located in this site will be crawled and analyzed. Besides tracking the URLs of these Web pages, the Site Crawler also records the depth (i.e. level) and length of each page. Depth represents the number of "hops" from the home page to the current page. For example, if we give the home page depth 1, then all pages which can be reached by an out-link of the home page are assigned depth 2. Length of a Web page is the number of characters in the Web page source file. The Site Crawler only keeps known types of Web pages, such as .htm, .html, .shtml, .php, etc. Handling other types of text and non-text files is a topic for future research.

Normally the *Site Crawler* crawls the top 1000 pages of a Web site, according to a breadth-first traversal starting from the home page. The number of pages to crawl (1000) is based on the observation after crawling 60 Web sites (identified in DMOZ subdirectories), that there is an average of 1000 pages up to and including depth equal to 4. For each Web site, the Site Crawler will stop crawling when either 1000 pages have been collected, or it has finished crawling depth 4, whichever comes first. Table 1 gives an example of samples of URLs, including depth and length, of the top 1000 pages from the Microsoft Corporation Web site.

After the URLs of the top 1000 Web pages are collected, the plain text must be extracted from these pages. Several packages are available for this purpose. Two of them, *HTML2TXT* [24] by Thomas Sahlin and *html2txt* [22] by Gerald Oskoboiny are compared against the text browser *Lynx* [10] and an *HTML Parser* developed by the authors. Our *HTML Parser* identifies all HTML tags in a Web page, removes dynamic content such as JavaScript, and

Order	URL	Depth	Length
1	http://www.microsoft.com	1	23472
2	http://www.microsoft.com/trainingandservices	2	38421
...	...	...	...
30	http://www.microsoft.com/downloads/search.asp	2	38401
31	http://www.microsoft.com/traincert	3	38421
...	...	...	...
514	http://www.microsoft.com/windows/default.asp	3	10689
515	http://www.microsoft.com/argentina/certificacion	4	17951
...	...	...	...
1000	http://www.microsoft.com/office/default.asp	4	16130

Table 1: Example of URLs, depth & length of top 1000 pages

keeps only plain text.

In order to test the individual performance of these four packages, 100 Web pages from DMOZ subdirectories were manually collected. There were no constraints such as size or type to these pages. Then the quality of the text parts generated by the four different packages is evaluated by the authors on a scale: *5-Excellent, 4-Good, 3-Satisfactory, 2-Poor,* or *1-Bad*. The quality is determined by considering both how much text is missing and how much non-text is included.

As indicated in Figure 2, *Lynx* is the most powerful package with an average score of 4.4, and hence in this work Lynx is used.

3 Narrative Text Extraction

The summary of the Web site will be created on the basis of the text extracted by Lynx. However, Web pages often contain isolated phrases, bullets or very short sentences, instead of a coherent narrative structure. Such text provides little foothold for a coherent and

No.	URL	P1	P2	P3	P4
1	http://www.mlnet.org	4	4	5	4
2	http://www.kdnuggets.com	3	4	4	3
3	http://www.rulequest.com	3	2	3	2
4	http://www.lcs.mit.edu	4	3	5	4
5	http://www.sun.com	3	4	4	3
6	http://java.sun.com	2	3	4	4
...	...	...	...	...	...
100	http://www.cs.berkeley.edu	4	3	5	3
	Average:	3.7	4.2	4.4	3.8

Scale: 5 --- Excellent
4 --- Good
3 --- Satisfactory
2 --- Poor
1 --- Bad

P1: *HTML2TXT* by Thomas Sahlin
P2: *html2txt* by Gerald Oskoboiny
P3: text browser *Lynx*
P4: *HTML Parser*

Figure 2: Performance comparison of different packages

meaningful summary [6], so our aim is to identify rules for determining which paragraphs should be considered for summarization and which ones should be discarded. This is achieved in two steps: First, criteria are defined for determining if a paragraph is long enough to be considered for analysis. Then, additional criteria are defined to classify long paragraphs into narrative or non-narrative. Only narrative paragraphs are used in summary generation. The criteria are defined automatically using supervised machine learning.

3.1 Long Paragraph Classification

During our experiments with the crawler, it is observed that some paragraphs are too short (in terms of number of words, number of characters, etc.) for summary generation, e.g., *This Web page is maintained by David Alex Lamb of Queen's University. Contact: dalamb@spamcop.net.*

Intuitively, whether a paragraph is long or short is determined by its length (i.e., the number of characters). However, two more features, number of words, and number of characters in all words, might also play key roles. In order to determine which feature is the

No.	Length	NumberOfWords	NumberOfChars	LONGSHORT
1	423	58	372	long
2	35	5	31	short
3	913	125	802	long
...	...	...	...	...
700	89	12	71	short

Table 2: Training data of C5.0 classifier *LONGSHORT*

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Size	2	2	2	2	2	2	2	2	2	2	2.0
Error(%)	5.7	5.7	11.4	4.3	2.9	4.3	4.3	7.1	2.9	10.0	5.9

Table 3: Cross-validation of C5.0 classifier *LONGSHORT*

most important, a total of 700 text paragraphs is extracted from 100 Web pages. Statistics of three attributes *Length, NumberOfWords* and *NumberOfChars* are recorded from each paragraph. *Length* is the number of all characters in the paragraph. *NumberOfWords* is the number of words in this paragraph, and *NumberOfChars* is the total number of characters in all words. Then each text paragraph is labelled as *long* or *short* manually. The decision tree learning program C5.0 [2] is used to construct a classifier, *LONGSHORT*, for this task.

The training set consists of 700 instances. Each instance consists of the values of three attributes and the associated class, as shown in Table 2. The resulting decision tree and its evaluation via ten-fold cross-validation are shown in Fig. 3. Among the 700 cases, there are 36 cases misclassified, leading to an error of 5.1%. The cross-validation of the classifier is listed in Table 3. The mean error rate 5.9% indicates the classification accuracy of this classifier.

Decision tree:

NumberOfWords < 20:

short (430)

NumberOfWords >= 20:

long (270/36)

Evaluation on training data (700 cases)

Decision Tree

Size	Errors	
2	36	(5.1%)
(a)	(b)	<<classified as
234	0	(a): class *long*
36	430	(b): class *short*

Figure 3: Decision tree of *LONGSHORT* and its evaluation

3.2 Narrative Paragraph Classification

Not all long paragraphs provide coherent information in terms of generating a meaningful summary. Intuitively, among the long paragraphs, narrative ones provide more coherent and meaningful content than non-narrative ones.

An example of a narrative paragraph is: *The users login file usually defines the command aliases and author identification (for the update history). Then one normally opens one or more CMZ files depending on the size of the project.*

An example of a non-narrative paragraph: ** ESTIMATE Professional (software project planning and estimation); * EssentialSET (documentation templates, process framework); * ISOplus (quality systems documentation)*

Informally, whether a paragraph is *narrative* or *non-narrative* is determined by the coherence of its text. We hypothesize that the frequencies of the part-of-speech tags of the words in the paragraph contain sufficient information to classify a paragraph as narrative. To test this hypothesis, a training set is generated as follows: First, 1000 Web pages are collected from DMOZ subdirectories, containing a total of 9763 text paragraphs, among which a total

of 3243 paragraphs were classified as long. Then, the part-of-speech tags for all words in these paragraphs are computed using a rule-based part-of-speech tagger [7].

The tagger accomplishes its task in two stages. First, every word is assigned its most likely tag in isolation. Unknown words are first assumed to be nouns (proper nouns if capitalized), and then cues based upon prefixes, suffixes, infixes, and adjacent word co-occurrence are used to change the guess of most likely tag. Second, contextual transformations are used to improve accuracy. For example, the tag of a word is changed from verb to noun if the previous word is tagged as a determiner. In this work, the list of 32 tags shown in Appendix A is used.

After part-of-speech tagging, the following attributes are extracted from each paragraph. Let n_i (i = 1, 2, ... , 32) be the number of occurrences of tag i, and S be the total number of tags (i.e. words) in the paragraph. Let P_i be the fraction of S, that n_i represents.

$$\begin{aligned} S &= \sum_{i=1}^{32} n_i \\ P_i &= n_i/S \quad (i = 1, 2, ..., 32) \end{aligned} \tag{1}$$

A total of 34 attributes are associated with each paragraph in the training set. The length of the paragraph in characters, and the length of the paragraph in words are added to the 32 attributes P_1, P_2, ..., P_{32}, as defined in equation 1. Then each paragraph is manually labelled as *narrative* or *non-narrative*. Finally, a C5.0 classifier *NARRATIVE* is trained on the training set of 3243 paragraphs, shown in Table 4.

The decision tree and its evaluation generated by the C5.0 program is presented in Figure 4. Among the 3242 cases, about 63.5% of them are following this rule: if the percentage of *Symbols* is less than 6.8%, and the percentage of *Preposition* is more than 5.2%, and the per-

No.	Length	Number	P_1	P_2	...	P_{32}	NARRATIVE
1	2010	201	0.040	0.020	...	0.003	narrative
2	1068	189	0.042	0.011	...	0.001	non-narrative
3	950	166	0.067	0.0	...	0.012	narrative
...	...	...	...	...	...	...	...
3243	572	108	0.020	0.049	...	0.020	non-narrative

Table 4: Training data of C5.0 classifier *NARRATIVE*

Decision Tree:

SYM > 0.068: *non-narrative* (354/14)
SYM <= 0.068:
:...IN <= 0.052: *non-narrative* (284/38)
  IN > 0.052:
  :...NNP <= 0.233: *narrative* (2058/90)
    NNP > 0.233:
    :...DT <= 0.075: *non-narrative* (236/72)
      DT > 0.075: *narrative* (210/46)

Evaluation on training data (3242 cases):

Decision Tree

Size	Errors	
5	260	(8.0%)
(a)	(b)	<< classified as
2232	124	(a): class *narrative*
136	750	(b): class *non-narrative*

Figure 4: Decision tree of *NARRATIVE* and its evaluation

centage of *Proper Singular Nouns* is less than 23.3%, then this paragraph is *narrative*. There are 260 cases misclassified, leading to an error of 8.0%. The cross-validation of the classifier *NARRATIVE* is listed in Table 5. The mean error rate 11.3% indicates the predictive accuracy of this classifier.

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Size	5	5	3	4	4	5	4	3	4	3	4.0
Error(%)	11.1	9.3	13.6	11.1	9.9	7.4	9.3	16.0	10.5	14.7	11.3

Table 5: Cross-validation of C5.0 classifier *NARRATIVE*

4 Key-Phrase Extraction

Traditionally, key-phrases (key-words and key-terms) are extracted from the document in order to generate a summary. Based on these phrases, the most significant sentences, which best describe the document, are retrieved.

Key-phrase extraction from a body of text relies on an evaluation of the importance of each phrase [9]. In terms of automatically summarizing a Web site, a phrase is considered as *key-phrase*, if and only if it occurs very frequently in the Web pages of the site, i.e., the total frequency is very high. This concept is different from the traditional TF/IDF measure [25], where a phrase within a given text is considered most important if it occurs frequently within the text, but infrequently in the larger collection of documents [9]. The validity of our *key-phrase* concept can be illustrated by a counterexample: suppose there is a page talking about Linux, and Linux occurs very frequently in this page, but rarely in the rest of the pages. Then Linux must be a good key-phrase based on TF/IDF. However, Linux is not good in describing the whole site.

In this work, a *key-phrase* can be either *key-word* or *key-term*. *Key-word* is a single word with very high frequency over the set of Web pages, and *key-term* is a two-word term with very high frequency.

As we discussed in the previous section, Web pages are quite different from traditional documents. The existence of *anchor text* and *special text* contributes much to the difference. *Anchor text* is the text of hyper links, and it "often provides more accurate descriptions of Web pages than the pages themselves" [8]. *Special text* includes title, headings and bold or italicized text. The assumption is that both anchor text and special text may play a key

No.	Word	f	fn	fa	fs
1	system	5450	4648	310	492
2	software	5313	3643	1092	578
3	product	2953	2033	668	252
...	...	...	...	...	...
5454	process	2294	1237	632	425

Table 6: Example of candidate key-words

role in describing important topics of Web pages. Therefore a supervised learning approach is applied to test this assumption.

4.1 Key-word Extraction

First we discuss how to produce decision tree rules for determining the key-words of a Web site. A data set of 5454 candidate key-words (at most 100 for each site) from 60 Web sites are collected. The sites are taken from DMOZ subdirectories. For each site, the frequency of each word in narrative text, anchor text and special text, is measured. Then the total frequency of each word over these three categories is computed, where the weight for each category is the same.

As it can be seen in Table 6, f is the total frequency of a candidate key-word, fn, fa, and fs are the frequencies of a candidate key-word in narrative text, anchor text and special text, respectively, hence $f = fn + fa + fs$. For example, the word *system* occurs 4648, 310 and 492 times in narrative text, anchor text and special text, respectively. This yields a total frequency of 5450. Moreover, it should be noted that 425 stop words (*a, about, above, across, after, again, against, all, almost, alone, along, already, also, although, always, among, an, and, ...*) [11] are discarded in this stage.

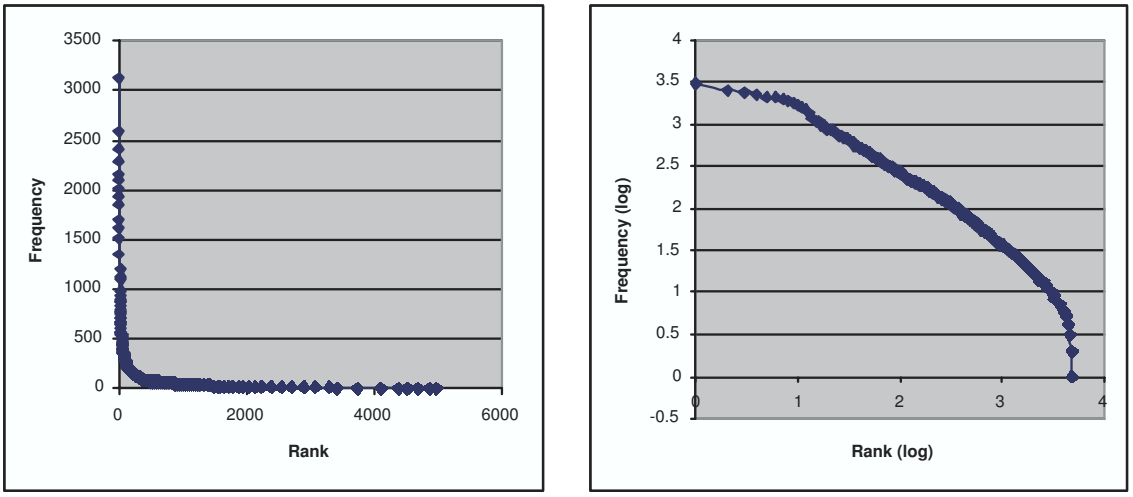

Figure 5: Rank-Frequency data and Zipf's Law

After all distinct words and their frequency statistics were recorded, a simple stemming process was applied to identify each singular noun and its plural form. Then the frequencies of the two are added. For example, *product* : 2100 and *products* : 460 yields *product* : 2560.

After this process, on the average there were about 5,100 different words (excluding stop words) within the text body of the top 1000 Web pages. The top 5 words often have a frequency of more than 2,000, which depends on the contents of Web pages, whereas the bottom ones may occur only twice or even once. Figure 5 shows that the rank and frequency statistics of these words fit Zipf's Law [18]. The bottom words are obviously not key-words, hence only those words whose frequency is more than 5% of the maximum frequency are kept as *candidate key-words*. This step eliminates about 98% of the original words, leaving about 102 candidate key-words per site. As a result, the top 100 candidate key-words are kept and nine features of each candidate key-word C_i are created, as shown in Table 7. The feature *Tag* was obtained by tagging candidate key-words with rule-based part-of-speech tagger [7].

No.	**Feature**	**Value**	**Meaning**
1	W	W_i=$f_i/\sum_{i=1}^{100} f_i$	Weight of candidate key-word C_i
2	R	R_i=$f_i/\max_{i=1}^{100} f_i$	Ratio of frequency to max frequency
3	WN	WN_i=$fn_i/\sum_{i=1}^{100} fn_i$	Weight in narrative text only
4	RN	RN_i=$fn_i/\max_{i=1}^{100} fn_i$	Ratio in *narrative* text only
5	WA	WA_i=$fa_i/\sum_{i=1}^{100} fa_i$	Weight in *anchor* text only
6	RA	RA_i=$fa_i/\max_{i=1}^{100} fa_i$	Ratio in *anchor* text only
7	WS	$WS_i = fs_i/\sum_{i=1}^{100} fs_i$	Weight in *special* text only
8	RS	$RS_i = fs_i/\max_{i=1}^{100} fs_i$	Ratio in *special* text only
9	Tag	$CC, CD, ..., WRB$	Part-of-speech tag (see Table 14)

Table 7: Feature list of candidate key-words

W	R	WN	RN	WA	RA	WS	RS	Tag	KEY-WORD
0.072	1.0	0.067	1.0	0.080	1.0	0.096	1.0	NN	key-word
0.047	0.651	0.055	0.824	0.017	0.214	0.039	0.403	NN	key-word
0.015	0.320	0.012	0.388	0.013	0.028	0.055	0.211	NN	key-word
...	...	...	...	...	...	...	...	...	...
0.010	0.136	0.007	0.104	0.026	0.323	0.005	0.051	VB	non-key-word

Table 8: Training data of C5.0 classifier *KEY-WORD*

Next, each candidate key-word is labelled manually as *key-word* or *non-key-word*. The criterion to determine if a candidate key-word is a true key-word is that a key-word provides important information which is related with the Web site. Based on frequency statistics and part-of-speech feature of these candidate key-words, a C5.0 classifier *KEY-WORD* is constructed as shown in Table 8.

The decision tree and its evaluation generated by the C5.0 program are presented in Figure 6. Among the total 5454 cases, 222 cases are misclassified, leading to an error of 4.1%. In the decision tree, about 35% of cases are following this rule: if R (defined as the ratio of a candidate key-word's frequency to the maximum frequency in Table 7) is less than or equal to 0.1, then this candidate key-word is a non-key-word. Another main stream of

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Size	22	20	20	30	23	18	20	27	20	20	22.0
Error(%)	4.0	5.1	5.5	4.4	4.0	5.1	5.1	5.9	5.5	4.0	4.9

Table 9: Cross-validation of C5.0 classifier *KEY-WORD*

cases follows the second rule: if R is greater than 0.1, and part-of-speech tag is NN (common singular nouns, see Table 14), and RA (ratio in anchor text) is less than or equal to 0.798, then the candidate key-word is a key-word. These cases cover 45% of the data set.

An interesting rule here is: if R is greater than 0.1, and part-of-speech tag is NNS (common plural nouns, see Table 14), then the candidate key-word should be classified as key-word. However, among these 138 cases, 50 were misclassified. This means that this training set was not effective in identifying common plural nouns, due to an insufficient number of such cases. The most important rule is: if R is greater than 0.1 and part-of-speech tag is NN (common singular nouns) or VBG (verb -ing, see Table 14), then WA (weight in anchor text), RA (ratio in anchor text) and/or WS (weight in special text) will determine if a candidate key-word should be classified as key-word or non-key-word. This demonstrates that our assumption is true, i.e., anchor text and special text do play important roles in determining key-words of a Web site. The cross-validation results of the classifier *KEY-WORD* is listed in Table 9. The mean error rate 4.9% indicates the predictive accuracy of this classifier.

4.2 Key-term Extraction

It is observed that terms which consist of two of the top 100 candidate key-words may exist with high frequency. Such a term could be good as part of the description of the Web

Decision Tree:

R <= 0.1: *non-key-word* (1908)
R > 0.1:
:...Tag in {CC,CD,EX,FW,IN,JJR,NNPS,NNP,
: PRP$,PRP,RBR,RBS,SYM,TO,VBZ,
: WDT,WP$,WP,WRB}: *non-key-word* (0)
  Tag = DT: *non-key-word* (2)
  Tag = JJS: *non-key-word* (4)
  Tag = JJ: *non-key-word* (350/6)
  Tag = MD: *non-key-word* (2)
  Tag = NNS: *key-word* (138/50)
  Tag = RB: *non-key-word* (18)
  Tag = UH: *non-key-word* (2)
  Tag = VBD: *non-key-word* (36)
  Tag = VBN: *non-key-word* (94/4)
  Tag = VBP: *non-key-word* (26)
  Tag = VB: *non-key-word* (292)
  Tag = NN:
  :...RA <= 0.798: *key-word* (2438/138)
  : RA > 0.798:
  : :...WA > 0.192: *non-key-word* (12)
  : WA <= 0.192:
  : :...RA <= 0.833: *non-key-word* (6)
  : RA > 0.833: *key-word* (52/14)
  Tag = VBG:
  :...WS <= 0.004: *non-key-word* (40/6)
  WS > 0.004:
  :...WS > 0.105: *non-key-word* (4)
  WS <= 0.105:
  :...R <= 0.121: *non-key-word* (4)
  R > 0.121: *key-word* (26/4)

Evaluation on **training data** (5454 cases):

Decision Tree

Size	Errors	
20	222	(4.1%)
(a)	(b)	<< classified as
2448	16	(a): class *key-word*
206	2784	(b): class *non-key-word*

Evaluation on **test data** (2718 cases):

Decision Tree

Size	Errors	
20	160	(5.9%)
(a)	(b)	<< classified as
1208	30	(a): class *key-word*
130	1350	(b): class *non-key-word*

Figure 6: Decision tree of *KEY-WORD* and its evaluation

site. For example, at the Software Engineering Institute Web site[1], the words *software* and *engineering* have frequency 7805 and 3430, respectively, and the term *software engineering* occurs 2492 times. Thus, a similar approach with *automatic key-word extraction* is developed to identify key-terms of the Web site.

The algorithm combines any two of the top 100 candidate key-words and searches for these terms in collocation over narrative text, anchor text and special text. Then these terms are sorted by frequency and the top 30 are kept as *candidate key-terms*. A C5.0 classifier *KEY-TERM* is constructed based on frequency statistics and tag features of 1360 candidate key-terms, which were extracted from 60 Web sites (collected from DMOZ subdirectories). The C5.0 classifier *KEY-TERM* is similar to the KEY-WORD classifier except that it has two part-of-speech tags *Tag1* and *Tag2*, one for each component word.

Once the decision tree rules for determining key-terms have been built, they are applied for automatic key-term extraction to the top 1000 Web pages of a Web site. The top 10 key-terms (ranked by total frequency) for each site are kept as part of the summary. Then the frequency of candidate key-words is reduced by subtracting the frequency of top 10 key-terms, which includes them. For example, the actual frequency of the words *software* and *engineering* above becomes $7805 - 2492 = 5313$ and $3430 - 2492 = 938$, respectively. Then candidate key-words of the Web site are classified into *key-word* or *non-key-word* by applying the KEY-TERM classifier shown in Figure 6. Finally, the top 25 key-words (ranked by frequency) are kept as part of the summary. It is observed that 40% to 70% of key-words and 20% to 50% of key-terms appear in the home page of a Web site.

[1] http://www.sei.cmu.edu

5 Key-Sentence Extraction

Once the key-words and key-terms are identified, the most significant sentences can be retrieved from all narrative paragraphs. Each sentence is assigned a significance factor or sentence weight. The top five sentences, ranked according to sentence weight, are chosen as part of the summary. In order to achieve this goal, we applied a modified version of the procedure in [9].

First, the sentences containing any of the list L of key-phrases, consisting of the top 25 key-words and top 10 key-terms identified previously, are selected.

Second, all clusters in each selected sentence S are identified. A *cluster* C is a sequence of consecutive words in the sentence for which the following is true: (1) the sequence starts and ends with a key-phrase in L, and (2) less than D non-key-phrases must separate any two neighboring key-phrases within the sentence. D is called the "distance cutoff", and we used a value of 2 as in [9]. Table 10 gives an example of clustering, where key-words, key-terms and clusters are listed.

Third, the weight of each cluster within S is computed. The maximum of these weights is taken as the sentence weight. A cluster weight is computed by adding the weights of all key-phrases within the cluster, and dividing this sum by the total number of words within the cluster [9]. The weight of key-phrase i is defined as $W_i = f_i / \sum_{i=1}^{100} f_i$, where f_i is the frequency of the key-phrase in the web site (Table 7). For example, the second cluster's weight in Table 10 is $(0.021 + 0.293 + 0.013)/5 = 0.065$.

However, division by the total number of words in the cluster decreases the weight too much when there are several key-phrases present together with a large number of non-

Candidate Sentence			
The Software Engineering Information Repository (SEIR) is a Web-based repository of information on software engineering practices that lead to improved organizational performance.			
Key-Phrase	**Weight**	**Cluster**	**Weight**
information	0.021	1. Software Engineering Information	0.157
software engineering	0.293	2. information on software engineering practices	0.109
practice	0.013	Sentence Weight: 0.157	

Table 10: Example of clustering

key-phrases. After some informal experimentation, the best cluster weighting (in terms of assigning the highest weight to the most informative sentences) is obtained by adding the weights of all key-phrases within the cluster, and dividing this sum by the total number of key-phrases within the cluster. Hence the second cluster's weight in Table 10 will now be $(0.021 + 0.293 + 0.013)/3 = 0.109$ and the first cluster's weight is $(0.293 + 0.021)/2 = 0.157$, therefore the sentence weight is 0.157.

The weights of all sentences in *narrative* text paragraphs are computed and the top five sentences ranked according to sentence weights are included in the summary as *key-sentences*. These key-sentences are expected to give a general idea of the contents of the Web site. Finally, a summary is formed consisting of the top 25 key-words, top 10 key-terms and top 5 key-sentences. Table 11 shows the generated summary of the Software Engineering Institute (SEI) Web site. This summary gives a brief description of SEI's mission, operator and various activities.

Part 1. Top 25 Key-words				
sei	system	software	cmu	product
component	information	process	architecture	organization
course	program	report	practice	project
method	design	institute	development	research
document	management	defense	technology	team

Part 2. Top 10 Key-terms				
software engineering	carnegie mellon	development center	software process	software architecture
maturity model	risk management	software development	process improvement	software system

Part 3. Top 5 Key-sentences
1. The Software Engineering Information Repository (SEIR) is a Web-based repository of information on software engineering practices that lead to improved organizational performance.
2. Because of its mission to improve the state of the practice of software engineering, the SEI encourages and otherwise facilitates collaboration activities between members of the software engineering community.
3. The SEI mission is to provide leadership in advancing the state of the practice of software engineering to improve the quality of systems that depend on software.
4. The Software Engineering Institute is operated by Carnegie Mellon University for the Department of Defense.
5. The Software Engineering Institute (SEI) sponsors, co-sponsors, and is otherwise involved in many events throughout the year.

Table 11: Summary of Software Engineering Institute Web site

6 Experiments and Evaluation

In order to measure the overall performance of our approach, four sets of experiments were performed. During these experiments, automatically generated summaries are compared with human-authored summaries, home page browsing and time-limited site browsing, to measure their performance in a specific task. Moreover, the performance of our approach is measured separately on academic Web sites and commercial ones, to test if there is any significant difference between them.

6.1 W3SS and DMOZ Summaries

From the DMOZ Open Directory Project, 20 manually constructed summaries were selected from four subdirectories. As listed in Table 12, sites 1-5 are in the *Software/Software Engineering*[2] subdirectory. Sites 6-10 are in the *Artificial Intelligence/Academic Departments*[3] subdirectory. Sites 11-15 are in *Major Companies/Publicly Traded*[4] subdirectory. And finally sites 16-20 are in *E-Commerce/Technology Vendors*[5] subdirectory. These sites were selected randomly and are of varying size and focus. Sites 1-10 are academic ones, with a focus on academic or non-profit research and/or development, whereas sites 11-20 are commercial ones, which deliver commercial products and/or services.

Our approach, W3SS (World Wide Web Site Summarization), is used to create summaries of these 20 Web sites. Each W3SS summary consists of the top 25 key-words, the top 10 key-terms and the top 5 key-sentences.

[2]http://dmoz.org/Computers/Software/Software_Engineering/

[3]http://dmoz.org/Computers/Artificial_Intelligence/Academic_Departments/

[4]http://dmoz.org/Business/Major_Companies/Publicly_Traded/

[5]http://dmoz.org/Business/E-Commerce/Technology_Vendors/

Subdirectory	Site URL
Software/ Software Engineering	1. http://case.ispras.ru 2. http://www.ifpug.org 3. http://www.mapfree.com/sbf 4. http://www.cs.queensu.ca/Software-Engineering 5. http://www.sei.cmu.edu
Artificial Intelligence/ Academic Departments	6. http://www.cs.ualberta.ca/~ai 7. http://www.ai.mit.edu 8. http://www.aiai.ed.ac.uk 9. http://www.ai.uga.edu 10. http://ai.uwaterloo.ca
Major Companies/ Publicly Traded	11. http://www.aircanada.ca 12. http://www.cisco.com 13. http://www.microsoft.com 14. http://www.nortelnetworks.com 15. http://www.oracle.com
E-Commerce/ Technology Vendors	16. http://www.adhesiontech.com 17. http://www.asti-solutions.com 18. http://www.commerceone.com 19. http://www.getgamma.com 20. http://www.rdmcorp.com

Table 12: URL list of the Web sites used in the experiments

6.2 Summarization Evaluation

There are two major types of summarization evaluations: *intrinsic* and *extrinsic* [16, 21]. Intrinsic evaluation compares automatically generated summaries against a gold standard (ideal summaries). Extrinsic evaluation measures the performance of automatically generated summaries in a particular task (e.g., classification). Extrinsic evaluation is also called task-based evaluation and it has become more and more popular recently [23]. In this work, extrinsic evaluation is used.

In extrinsic evaluation, the objective was to measure how informative W3SS summaries, DMOZ summaries, home page browsing and time-limited site browsing are in answering a set of questions (Appendix D) about the content of the Web site. Each question is meant

to have a well-defined answer, ideally explicitly stated in the summary, rather than being open-ended. Four groups of graduate students in Computer Science (5 in each group) with strong World Wide Web experience were asked to take the test as follows:

The first and second group was asked to read each W3SS and DMOZ summary, respectively and then answer the questions. The third group was asked to browse the home page of each of the 20 Web sites and answer the questions. The last group was asked to browse each Web site for at most 10 minutes (time-limited site browsing) and answer all questions. All answers were then graded in terms of their quality in a scale 0-20. The grades are tabulated in Appendix E.

Evaluation of W3SS Summaries The average score of all summaries over five subjects is 15.0. The average score of each W3SS summary over five subjects varies from 8.8 to 19.6, which means the quality of W3SS summaries varies from site to site. Summaries of Nortel Networks and Microsoft Corporation Web sites get the top two average scores 19.6 and 19.2, respectively. These two summaries give a brief but accurate description of the corporations. The summary of the Adhesion Technologies Web site gets the lowest average score 8.8, because it describes specific software products the company delivers but no general information about the company is available. The summary with the second lowest average score 10.0 corresponds to a site that contains archives of Software Engineering but the summary gives a description of specific software tools and cannot convey any sense of that this site is an information resource. However, the variance between the average scores of all summaries over five subjects is only 0.213, which shows that all subjects in this experiment evaluated W3SS summaries consistently.

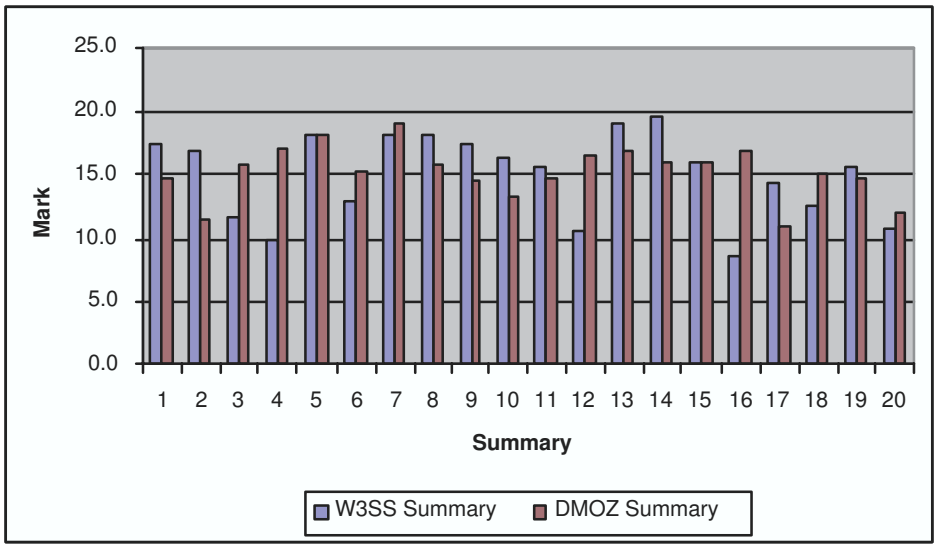

Figure 7: W3SS summaries vs. DMOZ summaries

Evaluation of DMOZ Summaries The average score of all summaries over all subjects is 15.3, hence the overall performance of DMOZ summaries is slightly better than that of W3SS ones (with an overall average 15.0). The average score of each DMOZ summary over five subjects varies from 11.0 to 19.2. However, the variance between the average scores of all DMOZ summaries over five subjects is 1.267, much larger than that of W3SS summaries.

As indicated in Figure 7, there are 11 Web sites whose W3SS summaries are better than DMOZ summaries, and 8 sites whose W3SS summaries are worse than DMOZ summaries. The remaining site has the same quality of W3SS and DMOZ summary.

Evaluation of home page browsing Since every subject was allowed to browse only the home page, there are a few very poor marks as low as 4.4 and 5.0. The average score of all home pages over five subjects is 12.7, which is less than 15.0 of W3SS summaries and 15.3 of DMOZ summaries.

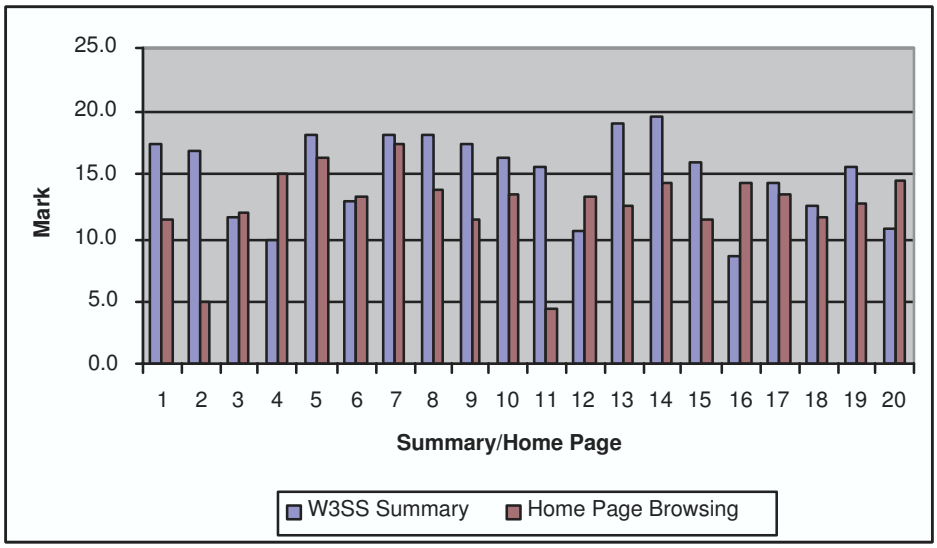

Figure 8: W3SS summaries vs. Home page browsing

As indicated in Figure 8, the performance of W3SS summaries is better than or the same as that of home page browsing. This experiment tells us that the home page alone is often not sufficiently informative, and that digging deeper into the site conveys more complete information about the site than the home page alone. In order to understand the site better, more browsing beyond the home page alone is needed.

Evaluation of time-limited site browsing In this test, every subject was allowed 10 minutes to browse each Web site, and look for the answers of all questions. For each site, the average score of all subjects varies from 7.0 to 20.0. This implies that either some Web sites were poorly designed, or there is too much non-text (e.g., flash)in top-level pages, which may confuse the user's understanding of the site. The average score of each site browsing over all subjects is 13.4, which is less than that of both W3SS and DMOZ summaries.

As indicated in Figure 9, the performance of W3SS summaries is generally better than

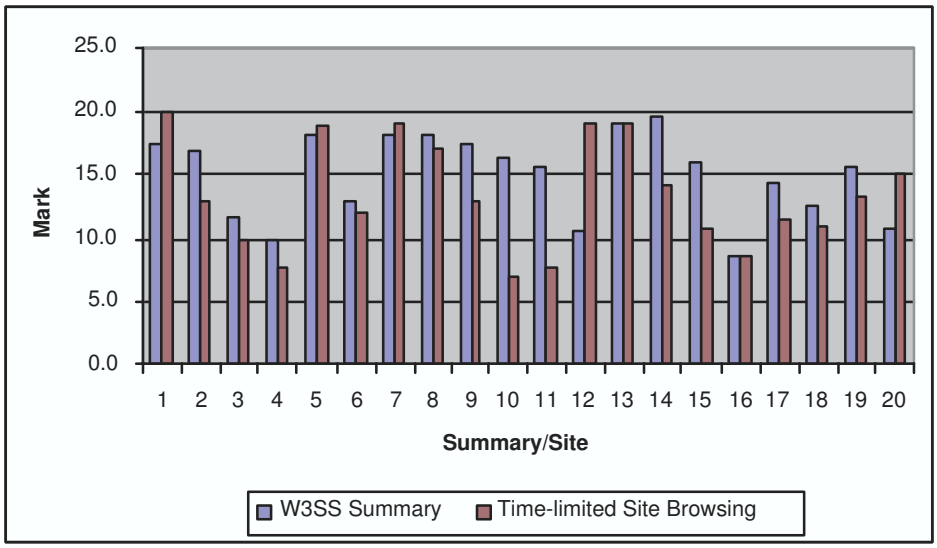

Figure 9: W3SS summaries vs. Time-limited site browsing

that of time-limited site browsing. This means it is not so easy to get a good understanding of the site's main contents by browsing within a limited time period. If the W3SS summary of a Web site is available, then the reader can know the site's main contents by viewing the summary without spending much time in browsing the site. This indicates that our approach of automatically creating summaries is potentially useful because it saves the reader much time.

To confirm the above intuitive conclusions, we perform a two-factor Analysis of Variance with replications on the raw scores from the above experiments. As shown in Table 13, there is no significant difference between our summaries and the human-authored summaries, and between home-page and time-limited site browsing. However, our summaries and the human-authored summaries are significantly better than home-page and time-limited site browsing.

Since the W3SS summaries are as informative as DMOZ summaries, they could be trans-

	W3SS	DMOZ	HPB
DMOZ	$F_{1,190} = 0.18$ $Pvalue = 0.67$		
HPB	$F_{1,190} = 17.42$ $Pvalue < 0.0001$	$F_{1,190} = 23.7$ $Pvalue < 0.0001$	
TLSB	$F_{1,190} = 6.13$ $Pvalue = 0.014$	$F_{1,190} = 8.88$ $Pvalue = 0.003$	$F_{1,190} = 1.62$ $Pvalue = 0.20$

Table 13: Pairwise ANOVA results for the four experiments. W3SS, DMOZ, HPB, TLSP is the performance of our summaries, the human-authored summaries, home-page browsing and time-limited site browsing.

formed into proper prose by human editors without browsing the Web site.

7 Conclusion and Discussion

In this work, we developed a new approach for generating summaries of web sites. Our approach applies machine learning and natural language processing techniques to extract and classify narrative paragraphs from the web site, from which key-phrases are then extracted. Key-phrases are in turn used to extract key-sentences from the narrative paragraphs that form the summary, together with the top key-phrases. We demonstrate that our summaries, although not in proper prose, are as informative as human-authored summaries, and significantly better than browsing the home page or the site for a limited time. Our approach should be easy to transform into proper prose by human editors without having to browse the web site. The performance of our method depends on the availability of sufficient narrative content in the web site, and the availability of explicit narrative statements describing the site.

Several issues need to be addressed to further improve the performance of our approach.

- Currently the top 1000 (or all pages between depth 1 and depth 4, inclusively) Web pages of a Web site are crawled for text extraction. Supervised learning may be used instead to determine the most appropriate number of pages to crawl. Appropriate attributes for this may include page depth, page length, or part-of-speech tags. Also during the current crawling process, only Web pages in ASCII coding are kept, while pages, which consist of binary contents such as movie clip, flash, are discarded. Such binary content may be a good indicator of the site's contents, hence it can be collected, analyzed and included as part of a summary.

- Currently the system (mainly written in Java) is running on a UNIX machine with 8 processors (400MHz UltraSPARC II) and 3GB Memory. In general, the amount of time required for the URL Extraction and Text Extraction steps depends on the throughput of the Internet connection. There is an average of 17000 paragraphs (including sentence segments) in the text parts of top 1000 Web pages. Long paragraph classification takes about one minute. Usually it takes 3 seconds for the part-of-speech tagger [7] (written in C) to tag a text file with around 100 words in this environment. So tagging an average of 5000 long paragraphs can last as long as 4 hours. It takes about three minutes to identify 1500 narrative paragraphs. Key-phrase extraction takes about 10 minutes, and key-sentence extraction needs about a minute. So more than 80% of computing time is spent in tagging long paragraphs.

- In the key-term extraction step, we simply combine any two of top 100 candidate key-words. More sophisticated methods, such as the C-value/NC-value method [12] will be considered to automatically recognize multi-word terms and then measure their

frequencies for the purpose of extracting key multi-word terms.

- Research in [8] indicates that assigning higher weight to anchor text may provide better results in search engines. Further research is required to determine appropriate weights for the key-words from different categories (plain text, anchor text and special text).

- Different subjects gave different marks to the same summary (or home page, site), indicating the potential presence of an "inter-rater reliability" [27] problem. Redesign of the evaluation process to reduce the inter-rater reliability problem is a topic for future research. Intrinsic evaluation should also be considered.

Acknowledgements We are thankful to Prof. Michael Shepherd for many valuable suggestions on this work, and to Jinghu Liu for suggesting the use of Lynx for text extraction from Web pages. The research has been supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Netscape 1998-2002. *DMOZ - Open Directory Project*. `http://dmoz.org`, last accessed on Oct. 9, 2002.

[2] RULEQUEST RESEARCH 2002. *C5.0: An Informal Tutorial*. `http://www.rulequest.com/see5-unix.html`, last accessed on Oct. 9, 2002.

[3] E. Amitay and C. Paris. Automatically summarising web sites - is there a way around it? In *ACM 9th International Conference on Information and Knowledge Management*, 2000.

[4] C. Aone, M.E. Okurowski, J. Gorlinsky, and B. Larsen. A scalable summarization system using robust nlp. In *Proceeding of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 66–73, 1997.

[5] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, Spain*, 1997.

[6] A. Berger and V. Mittal. Ocelot: a system for summarizing web pages. In *Proceedings of SIGIR*, pages 144–151, 2000.

[7] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, ACL*, 1992.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *7th International World Wide Web Conference*, 1998.

[9] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: Text summarization for web browsing on handheld devices. In *Proceedings of 10th International World-Wide Web Conference*, 2001.

[10] Internet Software Consortium. *Lynx: a World Wide Web (WWW) client for cursor-addressable, character-cell display devices*. `http://lynx.isc.org`, last accessed on Oct. 9, 2002.

[11] C. Fox. *Lexical analysis and stoplists, In W. Frakes and R. Baeza-Yates, editors, Information Retrieval: Data Structures & Algorithms*. Prentice Hall, Englewood Cliffs, NJ, chapter 7, pages 102130, 1992.

[12] K.T. Frantzi. Automatic recognition of multi-word terms. In *Ph.D. thesis, Manchester Metropolitan University, England*, 1998.

[13] T. Fukusima and M. Okumura. Text summarization challenge: Text summarization evaluation in japan. In *Proceedings of the NAACL'2001 Workshop on Automatic Summarization, Association for Computational Linguistics*, pages 40–48, 2001.

[14] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of SIGIR*, pages 121–128, 1999.

[15] J. Goldstein, V.O. Mittal, J.G. Carbonell, and J.P. Callan. Creating and evaluating multi-document sentence extract summaries. In *CIKM'00*, pages 165–172, 2000.

[16] S. Jones and J. Galliers. *Evaluating Natural Language Processing Systems: an Analysis and Review*. Springer, New York, 1996.

[17] S. Jones, S. Lundy, and G.W. Paynter. Interactive document summarisation using automatically extracted keyphrases. In *Proceedings of the 35th Hawaii International Conference on System Sciences*, 2002.

[18] Wentian Li. *Zipf's Law*. `http://linkage.rockefeller.edu/wli/zipf`, last accessed on Oct. 9, 2002.

[19] I. Mani. Recent developments in text summarization. In *CIKM'01*, pages 529–531, 2001.

[20] I. Mani, D. House, G. Klein, L. Hirschman, L. Orbst, T. Firmin, M. Chrzanowski, and B. Sundheim. The tipster summac text summarization evaluation. *Technical Report, MITRE, McLean, Virginia*, MTR98W0000138, October 1998.

[21] I. Mani and M. Maybury. *Advances in Automatic Text Summarization*. MIT Press, ISBN 0-262-13359-8, 1999.

[22] Gerald Oskoboiny. *html2txt*. `http://cgi.w3.org/cgi-bin/html2txt`, last accessed on Oct. 9, 2002.

[23] D.R. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Summarization Workshop*, 2000.

[24] Thomas Sahlin. *HTML2TXT*. `http://user.tninet.se/~jyc891w/software/html2txt/`, last accessed on Oct. 9, 2002.

[25] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

[26] IBM Research Laboratory Tokyo. *Automatic Text Summarization*. `http://www.trl.ibm.com/projects/langtran/abst_e.htm`, last accessed on Oct. 9, 2002.

[27] Colorado State University. *Writing Guide: Interrater Reliability*. `http://writing.colostate.edu/references/research/relval/com2a5.cfm`, last accessed on Oct. 9, 2002.

Tag	Meaning & Example	Tag	Meaning & Example
CC	conjunction (and, or)	RBR	adverb, comparative (faster)
CD	number (four, fourth)	RBS	adverb, superlative (fastest)
DT	determiner, general (a, the)	RB	adverb, general (fast)
EX	existential (there)	SYM	symbol or formula (US$500)
FW	foreign word (ante, de)	TO	infinitive marker (to)
IN	preposition (on, of)	UH	interjection (oh, yes, no)
JJR	adjective, comparative (lower)	VBD	verb, past tense (went)
JJS	adjective, superlative (lowest)	VBG	verb, -ing (going)
JJ	adjective, general (near)	VBN	verb, past participle (gone)
MD	modal auxiliary (might, will)	VBP	verb, (am, are)
NNPS	noun, proper plural (Americas)	VBZ	verb, -s (goes, is)
NNP	noun, proper singular (Canada)	VB	verb, base (go, be)
NNS	noun, common plural (cars)	WDT	det, wh- (what, which)
NN	noun, common singular (car)	WP$	pronoun, possessive (whose)
PRP$	pronoun, possessive (my, his)	WP	pronoun (who)
PRP	pronoun, personal (I, he)	WRB	adv, wh- (when, where, why)

Table 14: Rule-based part-of-speech tag list (from [7])

A Part-of-speech tag list

Table 14 shows the part-of-speech tags from [7] used in our system.

B DMOZ Summaries of the Web Sites

- 1. `http://case.ispras.ru/`

  Department for CASE Tools - Part of the Institute for System Programming of the Russian Academy of Sciences, this is a research and development organization focusing on Computer-Aided Software Engineering. Particular areas of interest are formal language processing, including compilers, development environments, visualization tools, reverse engineering, verification, repositories and Web portals.

- 2. http://www.ifpug.org/

  International Function Point User Group - A non-profit organization promoting the use of function point analysis and other software metrics.

- 3. http://www.mapfree.com/sbf/

  Software Build and Fix: Tips - Some tips to help programmers produce quality code.

- 4. http://www.cs.queensu.ca/Software-Engineering/

  Software Engineering Archives - World-Wide Web archives for USENET newsgroup comp.software-eng.

- 5. http://www.sei.cmu.edu/

  Software Engineering Institute (SEI) - SEI is a federal research center whose mission is to advance the state of the practice of software engineering to improve the quality of systems that depend on software. SEI accomplishes this mission by promoting the evolution of software engineering from an ad hoc, labor-intensive activity to a discipline that is well managed and supported by technology.

- 6. http://www.cs.ualberta.ca/~ai/

  AI Lab at University of Alberta - Information about their research, people, events, partners, courses, and more.

- 7. http://www.ai.mit.edu/

  AI Laboratory at MIT - The artificial intelligence laboratory at MIT.

- 8. http://www.aiai.ed.ac.uk/

  Artificial Intelligence Applications Institute - Information about the group, its people,

technologies, publications, events, clients, projects, employment opportunities.

- 9. `http://www.ai.uga.edu/`

  Artificial Intelligence Center - University of Georgia - Offers an inter- disciplinary masters degree in Artificial Intelligence and a bachelor degree in Cognitive Science. Strengths include logic programming, expert systems, neural nets, genetic algorithms, and natural language processing.

- 10. `http://ai.uwaterloo.ca/`

  Univ. of Waterloo Logic Programming and AI Group - Research in natural language processing, knowledge representation, machine translation, probabilistic reasoning, and planning.

- 11. `http://www.aircanada.ca/`

  Air Canada - Provides scheduled and chartered air transportation. (Nasdaq: ACNAF)

- 12. `http://www.cisco.com/`

  Cisco Systems, Inc. - Develops, manufactures, markets and supports high performance multiprotocol internetworking systems which link geographically dispersed local-area and wide-area networks, including wide range of multiprotocol routers, switches and dial access server products. (Nasdaq: CSCO).

- 13. `http://www.microsoft.com/`

  Microsoft Corporation - Designs, develops, manufactures, licenses, sells and supports a wide range of software products. (Nasdaq: MSFT).

- 14. `http://www.nortelnetworks.com/`

Nortel Networks - Telephony and IP-based data, wireline and wireless networking. (NYSE: NT).

- 15. http://www.oracle.com/

  Oracle Corporation - Designs, develops, markets and supports computer software products including database management and network products, applications development productivity tools and end-user applications, enabling the ability to retrieve, manipulate and control data stored on multiple computers, develop web-based client server applications, support operational requirements of on-line processing, decision support and data warehouse environments for high systems availability and performance, perform rapid querying and reporting, and multidimensional analysis of data. (Nasdaq: ORCL).

- 16. http://www.adhesiontech.com/

  Adhesion Technologies - A leader in providing integrated Internet solutions to better serve customers of established firms while simultaneously providing heightened revenue streams for core businesses.

- 17. http://www.asti-solutions.com/

  ASTi - Web order management system includes order entry, order processing, customer service, returns, credit card charges, and warehouse management, tools for site creation, and automated catalog import.

- 18. http://www.commerceone.com/

  Commerce One - Delivers business-to-business electronic commerce solutions to enter-

prise class organizations by dynamically linking buying and supplying organizations into real-time trading communities.

- 19. `http://www.getgamma.com/`

  Gamma - Shaping the landscape of the e-revolution. Gamma is a global IT solutions provider offering products and services to clients that are developing and operating e-commerce business models.

- 20. `http://www.rdmcorp.com/`

  RDM echeck - RDM is on the Financial Services Technology Consortium (FSTC) Electronic Check Project (www.echeck.org). This FSTC work has formed the basis for RDM's entry into the emerging electronic commerce business (e-check software).

C W3SS Summaries of the selected Web Sites

- 1. `http://case.ispras.ru/`

  Top 25 Key-words

  software, research, development, tool, technology, class, method, interface, language, command, engineering, option, prototype, process, object, character, window, figure, argument, element, text, list, number, image, server

  Top 10 Key-terms

  software engineering, reverse engineering, return value, class object, interface prototype, sdl mode, default value, value option, source code, interface type

Top 5 Key Sentences

(1) Department for Computer-Aided Software Engineering (CASE) Tools is part of the Institute for System Programming of the Russian Academy of Sciences.

(2) We are a research and development organization.

(3) Our expertise is in a broad range of technologies, which can be used to build powerful tools for software developers.

(4) We focus on formal language processing, including compilers, development environments, visualization tools, reverse engineering, verification, repositories and Web portals.

(5) Using Interface Editor and Interface Simulator from the toolkit, user is allowed to create interface prototypes and obtain sequence diagrams from them with Interface Simulator.

- 2. `http://www.ifpug.org/`

  Top 25 Key-words

  project, software, workshop, member, counting, isbsg, conference, box, information, membership, download, text, selection, cpm, article, newsletter, bylaw, course, measurement, process, organization, application, analysis, participant, group

  Top 10 Key-terms

  function point, academic affair, white paper, member search, vision plan, activity chart, target audience, workshop description, workshop number, workshop title

  Top 5 Key Sentences

  (1) The International Function Point Users Group (IFPUG) is a non-profit, member

governed organization.

(2) The mission of IFPUG is to be a recognized leader in promoting and encouraging the effective management of application software development and maintenance activities through the use of Function Point Analysis and other software measurement techniques.

(3) IFPUG maintains the Function Point Counting Practices Manual, the recognized industry standard for FPA.

(4) Also, through industry and academic relationships, IFPUG sponsors and supports projects for applied research on software measurement issues, and conducts studies in support of advancing the Function Point Counting Standards.

(5) The International Standards Organization (ISO) Task Group is working on the behalf of IFPUG members to advance Function Points as an international standard through ISO.

- 3. `http://www.mapfree.com/sbf/`

  Top 25 Key-words

  tcl, name, tk, widget, procedure, command, section, window, book, object, example, software, exercise, programmer, code, value, return, author, option, arg, character, argument, scripts, programming, version

  Top 10 Key-terms

  variable name, usage documentation, tk example, command window, tk page, tcl interpreter, command substitution, programming language, slave interpreter

  Top 5 Key Sentences

  (1) Tcl was meant to be portable and Tcl/Tk has been ported to versions of Microsoft

Windows and to the Macintosh.

(2) This company maintains a Tcl/Tk web site where you can download Tcl/Tk.

(3) Here's the table of contents with links to sample chapters covering introductory material in the Tcl and Tk sections as well as the chapters on regular expressions and the browser plugin.

(4) Almost two years back, the IEEE Computer Society published "Tcl/Tk for Programmers (with solved exercises that work with Unix and Windows)".

(5) A data type is essentially a record R together with procedures for manipulating objects of type R.

- 4. `http://www.cs.queensu.ca/Software-Engineering/`

  Top 25 Key-words

  software, tool, system, type, object, change, contact, entry, process, level, case, group, point, model, environment, project, version, development, management, ltd, information, example, design, product, name

  Top 10 Key-terms

  software engineering, case tool, tool gmbh, software process, software tool, software development, software corporation, code generation, function point, software product

  Top 5 Key Sentences

  (1) The Empirical Software Engineering Research Group at Bournemouth University maintains a bibliography on OO metrics, originally maintained by Robin Whitty of South Bank University.

  (2) There is a reference model of end-user services for software engineering environ-

ments (e.g., requirements, design, code, test, tracing, planning, publications, plus about 50 others) called the Project Support Environment Reference Model that was developed by the PSESWG (Project Support Environment Standards Working Group).

(3) Brad Myers (Brad.Myers@cs.cmu.edu) maintains a list of user interface software tools, which are tools that can help to create the user interface part of the software.

(4) The CASE vendor list and CASE tool list are both generated from the same (extremely simple) database from which I generate the monthly CASE tools FAQ on USENET.

(5) The quality of a software system is largely governed by the quality of the process used to develop and maintain the software.

- 5. `http://www.sei.cmu.edu/`

  Top 25 Key-words

  sei, system, software, cmu, product, component, information, process, architecture, organization, course, program, report, practice, project, method, design, institute, development, research, document, management, defense, technology, team

  Top 10 Key-terms

  software engineering, carnegie mellon, development center, software process, software architecture, maturity model, risk management, software development, process improvement, software system

  Top 5 Key Sentences

  (1) The Software Engineering Information Repository (SEIR) is a Web-based repository of information on software engineering practices that lead to improved organizational

performance.

(2) Because of its mission to improve the state of the practice of software engineering, the SEI encourages and otherwise facilitates collaboration activities between members of the software engineering community.

(3) The SEI mission is to provide leadership in advancing the state of the practice of software engineering to improve the quality of systems that depend on software.

(4) The Software Engineering Institute is operated by Carnegie Mellon University for the Department of Defense.

(5) The Software Engineering Institute (SEI) sponsors, co-sponsors, and is otherwise involved in many events throughout the year.

- 6. `http://www.cs.ualberta.ca/~ai/`

  Top 25 Key-words

  problem, system, search, research, model, method, information, algorithm, approach, agent, computer, image, constraint, technique, knowledge, network, motion, space, recognition, computation, graph, result, class, distribution, task

  Top 10 Key-terms

  artificial intelligence, belief net, problem solving, search space, learning algorithm, vision system, image process, computer game, computer vision, knowledge compilation

  Top 5 Key Sentences

  (1) Smodels has been successfully used to solve hard satisfiability problems and planning problems.

  (2) The research and development behind the system are built upon advanced methods

from the areas of Artificial Intelligence, Machine Learning and Data Mining, Computer Vision, Remote Sensing, and Silviculture.

(3) In practice, a "good" model of the problem is as important as a fast algorithm, as OR people have concluded that a "good" model of a problem is crucial to solve it efficiently.

(4) It has been recently recognized that both choosing the right solving algorithm and the right problem model are crucial for efficient problem solving.

(5) A solution to the information-overload problem is to create service agents which can gather and integrate information on behalf of a user.

- 7. `http://www.ai.mit.edu/`

  Top 25 Key-words

  system, image, lab, research, problem, model, paper, page, information, program, object, algorithm, computer, design, project, people, learning, result, method, language, group, example, number, function, point

  Top 10 Key-terms

  intelligence lab, artificial intelligence, lab operation, machine learning, graduate student, computer vision, object recognition, vision group, programming language, research project

  Top 5 Key Sentences

  (1) The Statistical AI Reading group (STAIR) meets weekly to host speakers and to read and discuss current and ongoing research in statistical methods in artificial intelligence and machine learning.

(2) Our intelligent machines must be able to learn from the world around them, and so we study better algorithms for learning from online data-bases and from sensory experiences.

(3) There has been a realization amongst many people at our lab that the keys to intelligence are self adapting perceptual systems, motor systems, and language related modules.

(4) In all areas we have had much success in the building of software and computer hardware systems.

(5) Our goal is to understand the nature of intelligence and to engineer systems that exhibit intelligence.

- 8. `http://www.aiai.ed.ac.uk/`

  Top 25 Key-words

  development, method, plan, activity, spar, research, process, model, aiai, issue, system, knowledge, project, task, agent, representation, constraint, type, description, level, information, detail, object, entity, domain

  Top 10 Key-terms

  artificial intelligence, implementation issue, attribute description, plan representation, object model, spar model, world state, enterprise ontology, core group, planning process

  Top 5 Key Sentences

  (1) AIAI was established in 1984 to encourage the development and take-up of artificial intelligence methods.

  (2) AIAI is a technology transfer organisation that promotes and expedites the appli-

cation of research on Artificial Intelligence for the benefit of industrial, commercial, government and academic clients.

(3) Also, AIAI is a tier 1 (enabling technology) research group within the ARPI with the O-Plan project.

(4) The work is drawing on results from previous projects at AIAI like Enterprise and O-Plan.

(5) AIAI and HSE staff worked together using the CommonKADS knowledge engineering methodology to model the task and capture the knowledge involved.

- 9. `http://www.ai.uga.edu/`

  Top 25 Key-words

  computer, term, programming, stream, prolog, list, center, goal, modeling, name, code, constraint, value, system, education, type, object, function, option, argument, module, program, integer, character, command

  Top 10 Key-terms

  artificial intelligence, genetic algorithm, computer modeling, expert system, debugger command, foreign resource, prolog flag, tcl command, input stream, object method

  Top 5 Key Sentences

  (1) The Artificial Intelligence Center is an interdepartmental research and instructional center within the Franklin College of Arts and Sciences of the University of Georgia.

  (2) Strengths include logic programming, expert systems, neural nets, genetic algorithms, and natural language processing.

  (3) The Artificial Intelligence Center also houses the undergraduate degree program in

Cognitive Science.

(4) The Artificial Intelligence Education Support Fund has been set up so that tax-deductible donations to The University of Georgia Foundation can be designated for the AI Center.

(5) Artificial intelligence is the computer modeling of intelligent behavior, including but not limited to modeling the human mind.

- 10. `http://ai.uwaterloo.ca/`

  Top 25 Key-words

  language, process, computer, vision, reasoning, translation, planning, knowledge, learning, class, frame, index, system, information, overview, package, program, type, rule, object, problem, field, parameter, chart, sentence

  Top 10 Key-terms

  computer vision, natural language, method summary, knowledge representation, probabilistic reasoning, machine translation, language model, computer science, machine learning, field summary

  Top 5 Key Sentences

  (1) The main activities of the group include research in: natural language processing, knowledge representation, machine learning, computer vision, probabilistic reasoning, machine translation, and planning.

  (2) Some ongoing projects include: statistical natural language modelling, reinforcement learning, and learning search control.

  (3) What is needed is a natural language generation system for the production of tai-

lored health-information and patient-education documents, that would, on demand, customize a "master document" to the needs of a particular individual.

(4) This is dependent on modeling whether the user has the knowledge the system is seeking, whether the user is willing to provide that knowledge and whether the user would be capable of understanding the request for information from the system.

(5) The project is centred at the University of Waterloo, with the additional participation of Graeme Hirst and his students at the University of Toronto, and Eduard Hovy from the Information Sciences Institute of the University of Southern California.

- 11. `http://www.aircanada.ca/`

  Top 25 Key-words

  service, canada, airline, flight, aeroplan, index, air, travel, aircraft, network, ticket, passenger, information, canadian, departure, carrier, business, arrival, type, market, mile, offer, world, need, customer

  Top 10 Key-terms

  air canada, privacy policy

  Top 5 Key Sentences

  (1) This site is designed specifically to provide convenient, quick access to information about Air Canada.

  (2) Welcome to the Air Canada Site Index! You can get to anywhere in our site from here, and it is easy to find the information you're looking for.

  (3) Air Canada will do everything practicable to ensure we provide the most accurate flight status information at all times.

(4) A series of statements, grouped under 13 themes, that clearly state the levels of service customers can expect from Air Canada and its regional carriers.

(5) On the ground and in the air, Air Canada's worldwide network of partners - airlines, hotels, car rental agencies and more - are here to help plan and enhance your next trip.

- 12. `http://www.cisco.com/`

  Top 25 Key-words

  cisco, network, command, service, router, interface, atm, support, voice, number, configuration, feature, address, mgc, access, step, configure, information, internet, solution, protocol, release, mpls, port, value

  Top 10 Key-terms

  cisco ios, privacy state, important notice, command mode, mpls traffic, command reference, configuration example, configuration guide, mpls ldp, configuration mode

  Top 5 Key Sentences

  (1) As the industry's only enterprise-wide, standards-based network architecture, Cisco AVVID provides the roadmap for combining your business and technology strategies into one cohesive model.

  (2) Cisco AVVID Network Infrastructure provides the baseline infrastructure that enables enterprises to design networks which scale to meet e-business demands.

  (3) Cisco AVVID Network Infrastructure delivers the e-business infrastructure and intelligent network services that are essential for rapid and seamless deployment of emerging technologies and new Internet business solutions.

  (4) Cisco recommends providers of network services who offer the highest levels of

quality and reliability in their network services.

(5) We work closely with your Cisco account manager, reseller, or channel partner to offer innovative, flexible financial services to Cisco customers and channel partners at competitive rates.

- 13. `http://www.microsoft.com/`

  Top 25 Key-words

  microsoft, window, product, server, support, service, information, office, software, search, page, business, right, training, solution, system, exam, resource, terms, technology, program, internet, download

  Top 10 Key-terms

  microsoft corporation, privacy statement, office xp, operating system, district court, free newsletter, microsoft window, microsoft office, microsoft project, microsoft certification

  Top 5 Key Sentences

  (1) Microsoft (Nasdaq "MSFT") is the worldwide leader in software, services and Internet technologies for personal and business computing.

  (2) The company offers a wide range of products and services designed to empower people through great software – any time, any place and on any device.

  (3) As the worldwide leader in software for personal and business computing, Microsoft strives to produce innovative products and services that meet our customers' evolving needs.

  (4) Whether you are a service provider, systems integrator, independent software ven-

dor, reseller or other type of technology provider, we depend on you to sell Microsoft solutions and products and to provide services for and build applications on the Microsoft platform.

(5) Microsoft offers comprehensive training courses to information technology (IT) professionals and developers who build, support, and implement solutions for practical, real-world situations using Microsoft products and technologies.

- 14. `http://www.nortelnetworks.com/`

  Top 25 Key-words

  service, solution, product, network, partner, industry, training, region, community, certification, care, country, documentation, store, business, internet, contact, system, advance, customer, location, portfolio, voice, search, copyright

  Top 10 Key-terms

  nortel network, customer support, corporate information, investor relation, media center, privacy statement

  Top 5 Key Sentences

  (1) Nortel Networks is an industry leader and innovator focused on transforming how the world communicates and exchanges information.

  (2) The company is supplying its service provider and enterprise customers with communications technology and infrastructure to enable value-added IP data, voice and multimedia services spanning Metro and Enterprise Networks, Wireless Networks and Optical Long Haul Networks.

  (3) As a global company, Nortel Networks does business in more than 150 countries.

(4) Service Activation is a key component of Nortel Networks service commerce solution.

(5) DISCLAIMER: The Nortel Networks Compatible Product(s) are verified as compatible in the laboratory environment with the indicated Nortel Networks Product(s) for a term of one (1) year from the date the Compatibility Test Audit/Evaluation was successfully completed until the Developer issues a new version of such Nortel Networks Compatible Product that incorporates new features or func- tionality or until Nortel Networks issues a new version of such Nortel Networks Product(s) listed in the compatibility certificate that incorporates new features or functionality, whichever occurs first.

- 15. `http://www.oracle.com/`

  Top 25 Key-words

  business, service, oracle, technology, suite, corporation, database, information, server, application, internet, customer, software, seminar, network, development, customer, enterprise, management, page, support, feature

  Top 10 Key-terms

  business suite, business network, oracle corporation, software development, application server, online service, internet seminar, oracle application, oracle database

  Top 5 Key Sentences

  (1) Oracle technology can be found in nearly every industry around the world and in the offices of 98 of the Fortune 100 companies.

  (2) Oracle is the first software company to develop and deploy 100% internet-enabled

enterprise software across its entire product line: database, business applications, and application development and decision support tools.

(3) Oracle is the world's leading supplier of software for information management, and the world's second largest independent software company.

(4) Oracle E-Business Suite has enabled more than 1,178 companies to become e-businesses, dramatically slashing costs and eliminating complexity.

(5) 65% of the Fortune 100 use Oracle for e-business.

- 16. `http://www.adhesiontech.com/`

  Top 25 Key-words

  adhesion, ea, solution, customer, technology, service, application, aggregation, ibm, client, partner, software, value, account, integration, contact, management, support, premise, information, career, product, bank, wealth, proact

  Top 10 Key-terms

  financial service, account aggregation, aggregation solution, hosting service, technology investment, aggregation software, aggregation investment, ea software, wealth management, financial institution

  Top 5 Key Sentences

  (1) Adhesion Technologies delivers Financial Service Providers (FSPs) an alternative to standard account aggregation services through its aggregation software called Enhanced Account Aggregation, or EA.

  (2) The result is a powerful aggregation software that financial services providers can use to better understand their customers, service their needs and generate a return on

their aggregation investment.

(3) Adhesion Technologies is a BroadVision Premier ASP Partner in the Financial Services Sector in North America and Europe.

(4) Adhesion's EA (Enhanced Account Aggregation) solution allows financial services providers to fully own and control their confidential customer data and to integrate that data with wealth management tools to generate a return on their technology investment.

(5) Using Adhesion's Enhanced Account Aggregation (EA), FSPs can deliver a private-label, personalized wealth management experience to their customers.

- 17. `http://www.asti-solutions.com/`

  Top 25 Key-words

  sale, service, company, operation, software, support, retail, contact, training, order, store, client, inventory, management, technology, system, marketer, website, need, product, catalog, solution, business, function, feature

  Top 10 Key-terms

  accessing information, applied system, dmms plus, classroom training, setup training, retail industry, support training, direct marketing

  Top 5 Key Sentences

  (1) Efficient and powerful software for direct marketers that brings success to e-tail, retail, and catalog operations.

  (2) Applied Systems Technologies, Inc is dedicated to the highest quality software and support in our industry.

(3) Our services include software maintenance and consulting services in all aspects of the direct marketing and retail industries.

(4) Whether you are aggressively researching a better software system for your business, or you simply feel a need to broaden your knowledge about software for direct marketing, you should start a relationship with ASTi.

(5) Whether you take 100 orders per day or 500,000 orders per day DMMS Plus can be customized to meet your needs.

- 18. `http://www.commerceone.com/`

  Top 25 Key-words

  commerce, solution, application, marketplace, business, internet, information, service, trading, process, platform, risk, integration, customer, software, network, communication, management, statement, consulting, release, suite, supplier, partner

  Top 10 Key-terms

  privacy information, global trading, marketplace solution, trading partner, risk factor, marketplace company, business process, commerce solution, global service, procurement solution

  Top 5 Key Sentences

  (1) Commerce One delivers the only e-commerce application suite built on an integration and process management platform that links and orchestrates enterprise legacy applications to deliver complete visibility and better communication across business units and trading partners.

  (2) Through its software, services and Global Trading Web business community, Com-

merce One enables worldwide commerce on the Internet.

(3) Commerce One.net is a web-based network for trading partners, providing Business Internet services to buyers and suppliers.

(4) Commerce One Global Services provides a full range of consulting services.

(5) Commerce One Global Services is dedicated to helping our customers achieve maximum results from Commerce One solutions.

- 19. `http://www.getgamma.com/`

  Top 25 Key-words

  infoshuttle, gamma, section, technologies, application, solution, service, initiative, development, business, network, system, object, client, test, environment, island, enterprise, company, system, certification, software, server, product, exam

  Top 10 Key-terms

  gamma enterprise, weblogic server, business application, course outline, application server, infoshuttle move, study guide, gamma privacy, enterprise component, business infrastructure

  Top 5 Key Sentences

  (1) Matthew Minkovsky is the founder of Gamma Enterprise Technologies, Inc.

  (2) Gamma provides software solutions to create, deploy and optimize critical business applications, making them run more efficiently and connecting them to partners and markets via the Internet.

  (3) Today, Gamma continues to help companies roll out reliable and scalable infrastructures for large e-business initiatives.

(4) Gamma provides complementary application and network performance engagements to assure that the end product has no performance bottlenecks.

(5) Our focus is helping companies that are adopting open standards such as J2EE, SOAP, and XML lower the cost of ownership of their business applications while using the Internet to provide a way to link to their trading partners.

- 20. `http://www.rdmcorp.com/`

  Top 25 Key-words

  rdm, scanner, image, links, check, document, micr, brochure, acrobat, checksheet, training, terminal, click, product, processing, bill, information, accessories, offer, control, market, roducts, service, ocr, quality

  Top 10 Key-terms

  payment archive, electronic check, check printer, micr test, rdm check, rdm iqx, bill stub, micr verifier, image test, bill payment

  Top 5 Key Sentences

  (1) RDM is a market leader in the check processing industry.

  (2) Many check printers have asked RDM how they could test checks that were printed not just on the bottom of a cut-sheet, but rather on the top or middle part of the 8 1/2" x 11" document.

  (3) RDM also offers National Training Programs to their ECC partners to help further educate the industry on the benefits of electronic check conversion with image.

  (4) RDM's PA Service was developed to house images that are captured through the RDM EC5000i scanner.

(5) RDM offers a number of different products that are used to verify that the MICR and OCR line on printed checks meet bank standards.

D Sample Questions for Extrinsic Evaluation

Questions

- 1. Is the purpose of the site to

  a. describe an academic organization

  b. describe a commercial company

  c. describe an information resource

  d. others, please state

- 1A. If the answer of Question 1 is a or b, then

  (1) Can you find the official name of the entity behind the site?

  a. yes, circle the name of the entity and label Answer 1A(1)

  b. no

  (2) Is the entity part of an organization?

  a. yes, circle the name of the organization and label Answer 1A(2)

  b. can't tell

- 2. What topic are the main contents of this site talking about?

  a. software engineering

  b. artificial intelligence

  c. publicly traded companies

d. e-commerce companies

e. others, please state

- 3. What's the main activity of the entity behind the site?

  a. conducting academic research or development

  b. delivering commercial products or services

  c. presenting some archives

  d. others, please state

- 4. Can you find the mission (focus, goal, interest, strength, etc.) of the entity behind this site?

  a. yes, circle the mission and label Answer 4

  b. no

- 1B. If the answer of Question 1 is a or b, then give a two-line description of the activity of the entity behind the site (no new text required, just highlight the sentences selected).

- 1C. If the answer of Question 1 is c, then give a two-line description of the contents of the site (no new text required, just highlight the sentences selected).

Marking scale

- Q1. 5 points: If the answer is a or b, then 2 points for this question plus 2 points for Q1A and 1 point for Q1B. If the answer is c, then 4 points for this question plus 1 point for Q1C.

- Q2. 5 points
- Q3. 5 points
- Q4. 5 points

Total points: 20

E Grades assigned in extrinsic evaluations

Tables 15, 16, 17, 18 show the grades assigned to the answers to the question list based on each of the four methods (inspection of summaries generated by our method, inspection of the human-authored summaries, homepage browsing and time-limited site browsing) respectively. Rows correspond to different web sites, while columns correspond to human evaluators.

W3SS	Mark 1	Mark 2	Mark 3	Mark 4	Mark 5	Total	Average
1	17	19	19	13	19	87	17.4
2	19	14	19	13	19	84	16.8
3	10	19	10	15	5	59	11.8
4	5	10	10	6	19	50	10.0
5	18	20	20	15	18	91	18.2
6	13	15	12	15	10	65	13.0
7	18	18	19	17	18	90	18.0
8	20	20	19	20	11	90	18.0
9	15	15	19	19	19	87	17.4
10	15	16	16	20	15	82	16.4
11	19	14	15	11	19	78	15.6
12	18	7	5	13	10	53	10.6
13	19	20	19	18	20	96	19.2
14	19	20	20	19	20	98	19.6
15	13	19	19	19	10	80	16.0
16	14	9	6	9	6	44	8.8
17	14	15	15	14	14	72	14.4
18	18	10	10	15	10	63	12.6
19	14	15	20	14	15	78	15.6
20	14	10	10	10	10	54	10.8
Average	15.6	15.3	15.1	14.8	14.4	75.1	15.0

Table 15: Performance of W3SS summaries

DMOZ	Mark 1	Mark 2	Mark 3	Mark 4	Mark 5	Total	Average
1	15	19	20	18	2	74	14.8
2	8	14	10	15	10	57	11.4
3	15	10	20	19	15	79	15.8
4	20	20	16	15	14	85	17.0
5	20	20	15	20	15	90	18.0
6	14	14	11	19	19	77	15.4
7	20	19	19	19	19	96	19.2
8	14	14	13	19	19	79	15.8
9	19	15	10	14	15	73	14.6
10	9	15	9	14	20	67	13.4
11	14	15	11	14	20	74	14.8
12	14	15	14	20	20	83	16.6
13	14	15	15	20	20	84	16.8
14	14	9	19	19	19	80	16.0
15	15	10	15	20	20	80	16.0
16	19	15	20	15	15	84	16.8
17	15	15	6	9	10	55	11.0
18	10	15	15	20	15	75	15.0
19	15	10	15	19	15	74	14.8
20	20	6	15	14	5	60	12.0
Average	15.2	14.3	14.4	17.1	15.4	76.3	15.3

Table 16: Performance of DMOZ summaries

Homepage	Mark 1	Mark 2	Mark 3	Mark 4	Mark 5	Total	Average
1	10	13	8	12	15	58	11.6
2	5	7	3	5	5	25	5.0
3	12	11	15	10	12	60	12.0
4	15	15	15	15	15	75	15.0
5	20	15	13	18	16	82	16.4
6	12	15	13	11	15	66	13.2
7	16	15	18	18	20	87	17.4
8	14	12	13	15	15	69	13.8
9	15	10	8	15	10	58	11.6
10	15	10	13	15	15	68	13.6
11	5	3	4	5	5	22	4.4
12	15	15	13	15	9	67	13.4
13	10	12	14	12	15	63	12.6
14	15	15	14	13	15	72	14.4
15	15	10	11	11	10	57	11.4
16	12	15	14	16	15	72	14.4
17	10	15	13	15	15	68	13.6
18	10	14	13	12	10	59	11.8
19	13	11	14	11	15	64	12.8
20	14	15	14	15	15	73	14.6
Average	12.7	12.4	12.2	13.0	13.1	63.3	12.7

Table 17: Performance of home page browsing

Site	Mark 1	Mark 2	Mark 3	Mark 4	Mark 5	Total	Average
1	20	20	20	20	20	100	20.0
2	15	11	12	12	15	65	13.0
3	10	7	10	11	12	50	10.0
4	10	7	8	9	5	39	7.8
5	20	16	20	18	20	94	18.8
6	10	15	15	9	11	60	12.0
7	20	15	20	20	20	95	19.0
8	10	15	20	20	20	85	17.0
9	10	15	15	10	15	65	13.0
10	10	5	5	5	10	35	7.0
11	7	10	5	8	9	39	7.8
12	20	20	20	15	20	95	19.0
13	20	20	20	15	20	95	19.0
14	15	15	15	10	16	71	14.2
15	10	10	10	9	15	54	10.8
16	10	5	9	15	5	44	8.8
17	15	15	8	10	10	58	11.6
18	15	10	10	10	10	55	11.0
19	15	12	14	15	11	67	13.4
20	15	15	15	15	15	75	15.0
Average	13.9	12.9	13.6	12.8	14.0	67.1	13.4

Table 18: Performance of time-limited site browsing