

Automatically Summarising Web Sites - Is There A Way Around It?

Einat Amitay

einat@ics.mq.edu.au

Division of Information and Communication Sciences
Macquarie University
NSW 2109 Australia

Cécile Paris

Cecile.Paris@cmis.csiro.au

CSIRO Mathematical & Information Sciences
Locked bag 17
North-Ryde
NSW 1670 Australia

ABSTRACT

The challenge of automatically summarising Web pages and sites is a great one. However, currently there is no solution which offers an easy way to produce unbiased, coherent, and content-full summaries of Web sites. In this work we suggest a new approach, which relies on the structure of hypertext and the way people describe information in it. As a proof-of-concept, we applied our approach to the problem of tailoring coherent snippets for search results. In this paper we describe the approach as it is implemented in our system, InCommonSense, and present results from a large scale evaluation of the snippets produced. We conclude by suggesting other applications that could make use of this technique.

Keywords

Information Retrieval from links, Web site summarisation.

1. INTRODUCTION

Being able to automatically and coherently describe the content of Web Sites has many advantages: We could present these summaries as search results, or archive them in a Web directory. We could also use the summaries to tailor short personalised bookmarks. There are currently several ways to achieve this goal: Some are fully automatic and use statistical/linguistic summarisation techniques, while others rely on Meta tags inserted by the authors of the Web page, or on human authored summaries that are prepared by editors of directories like the Open Directory Project [33]. There is currently no commercial search engine that uses the former technique (i.e., automatic statistical/linguistic summarisation techniques).

Permission to make digital or hard copies of part or all of this work or personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM 2000, McLean, VA USA

© ACM 2000 1-58113-320-0/00/11...\$5.00

This paper proposes a new fully automatic pseudo-summarisation technique. We use the term 'pseudo' because some may not agree that this technique could be put under the same umbrella as other, automatic, summarisation approaches. This new approach is unique since it does not analyse the document at all. Instead, it re-uses Web authoring conventions, understanding how useful information can be extracted from hypertext layout and language structure.

We chose to demonstrate the strength of this approach by applying the technique to generate short coherent textual snippets presented to the user with search engine results. This is only one application, and it is presented here as a proof-of-concept.

This paper will first outline the basic ideas behind the new approach. Then it will go on to describe an application, InCommonSense, that makes use of the technique to present Web search engine results with coherent textual summaries. Results from evaluating the output of InCommonSense against current state-of-the-art search engine output will also be presented. We will then conclude the paper by suggesting other applications for such an approach and state where our research is heading.

2. AUTOMATIC TEXT SUMMARISATION TECHNIQUES

Since we are interested in providing appropriate summaries describing the content of Web documents, it is useful to look at the state-of-the-art techniques in text summarisation. To be able to address questions like 'why use a new solution when there are many ready made summarisation systems out there?', we would like to bring forth the shortcomings of the research done in the field of text. In this short survey we will introduce the basic research directions in the field of text summarisation. The work surveyed here is of the kind that can address the special limitations of summarising documents on the Web, namely:

1. There is no restricted domain and "anything goes" on the Web, from journal articles to lists of links.
2. There is no syntactic markup and only physical properties such as paragraph and title/headers are marked, assuming the author of the document inserted them where appropriate.

3. In many cases the documents contain only a few words or sentence fractions which might not be considered cohesive text [2].
4. Summarisation technique used for online applications should require almost no pre-processing or training. It should also be completed as fast as possible.
5. The information on the Web might include images and other non-textual entities (e.g. audio, video etc.) which can not be identified as text but should be reported all the same.

In the automatic text summarisation research which addresses some of these issues there are three major trends. The first is the paragraph based summarisation ([11] [24] [20] and [1]), which tries to identify a single paragraph or text segment that addresses a single topic in the text. This method is usually employed when the text is long and there is more than one topic discussed, so that there is a need for supplying the user with an excerpt from the text. For the purpose of describing Web documents, it is not obvious that Web pages are structured in a way that a single paragraph can be retrieved. In the case of hypertext there is a tendency to segment the text into topics which are presented on different nodes and thus a single document might be too short for using this technique. The other problem is of course the use of graphics on the Web which such a technique is not able to summarise, or even to give some sort of indication as to the content of an image.

The second trend is the sentence-based abstraction technique. This technique relies on repetitions of terms and phrases in the text and tries to extract the most salient sentences or key-phrases and assemble them together ([23] [27] [6] [18] [28] [8] [26]). This technique assumes that a summary does not have to be cohesive, but that it should represent as much information from the text as possible. There are many researchers studying this field, and some of them use heuristics such as preferring terms which appear in titles or headers, preferring some terms which appear in bold or italic characters, etc. This technique might be more useful for summarising Web documents since it can make use of small portions of information, although it still can not handle graphic content.

The third and last trend is the discourse model based extraction and summarisation ([15] [7] [19] [16] [17] [25]). This technique is using natural language cues in the text such as lexical choice in the text, order of words, proper names identification, reiterations, synonymy, anaphora, lists of predefined cue phrases, connectives, etc. Again, this technique assembles sentences from the text based on these heuristics, and the resulting summary is usually a collage of facts and clues about what might be found in the text. Since some of the discourse features are domain specific, the systems in this field of research are tuned to one domain of text and to specific genres of writing. This fact makes the implementation of such systems for wide use on the Web very difficult.

The research interest in the field of text summarisation has been increasing in the last few years. Although there is not yet a solution to problems such as maintaining cohesion and coherence in the summaries, there is a lot of progress in the gathering of salient information from the text.

In this paper we would like to suggest an intermediary approach. An approach that takes into account the fact that, on the Web,

there are already many summaries of pages that people insert as annotated links, recommendation lists, and personal scribbles in Web pages. This approach is guided by several basic assumptions:

- i. People describe and comment (annotate) on other pages on the Web.
- ii. There are language patterns in the way people annotate (describe and comment) on the Web.
- iii. Many of these annotations can be automatically retrieved through understanding of the language patterns used and the way they were written in HTML.
- iv. Through experiments to assess the quality of the annotations, it is possible to characterise what makes a good/bad description/summary of a Web document, and to design a fast and automatic test for quality.

3. THE INCOMMONSENSE SYSTEM

The InCommonSense is a system that takes advantage of the paragraph convention found in Web hypertext [4]. It automatically extracts annotations, notes, descriptions and other scribbles that people write about other Web pages. InCommonSense was designed as a proof-of-concept of our approach. It is a system that tailors coherent snippets to search engine results. It is not a search engine by itself. Instead we designed a "leach" that sits on Web search engines of a certain kind. The kind of engines we chose for the purpose of our application are those that are able to process queries of the type "link:URL" (i.e., "tell me who links to this URL?").

3.1 Using paragraph writing conventions

The underlying approach in collecting information from arbitrary Web pages is that there is a pattern in the way people describe and link to other documents. This pattern is found in the way people write and annotate in hypertext. On the Web, there are different patterns of linking within the limits of a paragraph. These patterns can generally be viewed as four distinct patterns, as shown in Figure 1.

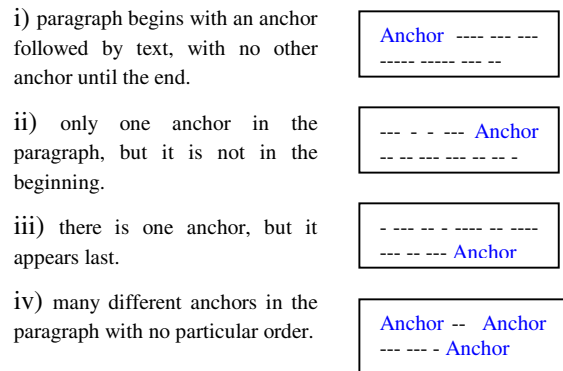


Fig. 1. Four paragraph-link patterns found on the Web.

In several experiments conducted with over 250,000 documents, the first pattern, a paragraph beginning with an anchor followed by text, was found to be useful for predicting the topic of the

linked document [4]. Since this form of writing is practised by many users and is becoming more widespread everyday, it can be efficiently and automatically manipulated. This is the pattern we are making use of in our system.

3.2 Overview of the system

The InCommonSense system takes a Web document A and looks for other documents D_i that point to it. Then the system looks for patterns of the type (i) in Figure 1 (anchor followed by text) in D_i that point to A, as these are likely descriptions of A. The relations between the documents found by the system is shown in Figure 2.

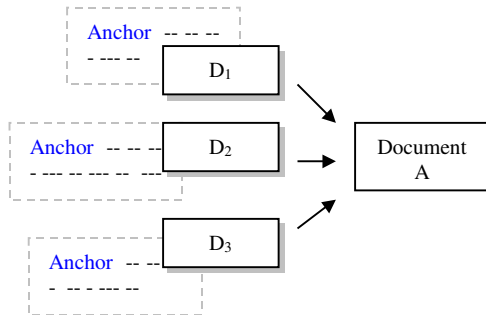


Fig. 2. Relations between documents as detected by the InCommonSense system.

InCommonSense collects information about specific documents by querying Web search engines for links to the document specified (query of the type - "link:URL"). Then the system fetches the pages and analyses them for paragraph markup cues (any markup line break). InCommonSense looks for segments of text that have empty spaces before and after the text, and segments of text that are marked as different entities (list items, table cells, text indents, etc.). This analysis helps the system detect visual cues like paragraph shapes and bulleted lists that readers might consider to be individual or standalone paragraphs of text. If the anchor linking to the required page is followed by text, as in paragraph type (i) in Figure 1, then both text and anchor are retrieved. The system also allows a determiner (e.g., *The, This, A*, etc.) to precede the anchor position, which was found to be a useful pattern in previous work[2][3]. For example, the following snippets of text (Fig. 3) relate to the document found in <http://www.geocities.com/CapitolHill/6228/> (titled - "Elections USA").

Currently, InCommonSense processes up to 220 documents relating to a single URL in one run (using Google[30], HotBot [31], AltaVista [29] and Infoseek [32]). There is no limit to the number of documents processed except for the limits that the commercial search engines pose: The Web is reflected through the search engines used by InCommonSense, which means that the documents processed are only the ones that are detected by the search engines.

1. [Anchor](#) -- "all the elections, all the results, all the time..."
source: <http://www.library.vanderbilt.edu/central/staff/fdtf.html>
2. [Anchor](#) (GeoCities) "Nonpartisan site, since 1997 bringing you the latest news from the campaign trail."
source: <http://www.igs.berkeley.edu:8880/library/agpp.html>
3. [Anchor](#) All the elections, all the results, all the time. Resource Type: Periodical
source: <http://galaxy.einet.net/galaxy/Government/United-States-Government/Politics/Elections.html>
4. [Anchor](#) An excellent site for current election news and analysis
source: <http://www.academicinfo.net/polisci.html>
5. [Anchor](#) : Geocities site aiming to become "the most objective site on the Web"
source: <http://www.aph.gov.au/library/intguide/pol/polelect.htm>
6. [Anchor](#) -- All big elections, all over the nation, all the time.
source: <http://www.clemens.org/pols.htm>
7. [Anchor](#) - Updated news stories on state and local races
source: <http://members.home.net/philvalentine/>
8. [Anchor](#) Provides analysis on campaigns, politics, interesting articles and research. <http://www.geocities.com/CapitolHill/6228/> [Visual Content:
source: <http://www.studyweb.com/links/1027.html>
9. [Anchor](#) Elections USA turns viewers into informed voters. Here you can find the latest news, polls, election returns, books on political candidates, and much more. Be sure to check this site out! <http://www.geocities.com/CapitolHill/6228/> [Visual Content:
source: <http://www.studyweb.com/links/1017.html>
10. [Anchor](#) Site useful for political news, with links to conservative and liberal web sites.
source: <http://library.uncg.edu/depts/ref/qil/politics.htm>
11. [Anchor](#) is an independent site that strives to offer a great deal of objective information. Includes a special section on the Clinton-Giuliani Senate race in New York, along with current poll data and a big list of links.
source: <http://www.portland.com/newslinks.shtml>

Fig. 3. Snippets of text that InCommonSense assumes to be related to the URL - <http://www.geocities.com/CapitolHill/6228/> (titled - "Elections USA").

4. CHOOSING THE BEST DESCRIPTION FROM THE RETRIEVED SNIPPETS

The InCommonSense system usually finds more than one description for each Web page, as illustrated in Figure 3. Since this is the case, to be useful, the system needs to be able to predict which of the descriptions is a better candidate for describing a given Web page. To this end, a filtering mechanism was designed to allow for sifting through all the snippets originally retrieved by the system automatically, choosing a single description for each Web page (or URL). This filter was designed to capture an understanding of users' preferences with respect to various online textual descriptions of Web documents. This understanding of users' preferences was achieved by designing a large scale online experiment.

4.1 Online descriptions and people's preferences

The goal of the experiment was to determine the descriptive value of the snippets collected with InCommonSense for the purpose of building a filter for identifying good descriptions in the data. The

experiment was designed to answer the need for understanding what people would prefer to see as a search result in terms of textual description of a Web page found by a search engine. The results were used to determine which language or textual features would be useful for automatically predicting good and bad descriptions. These features are used in the filtering mechanism component of the system.

First, it should be noted that this experiment does not attempt to model the whole Web, and each textual snippet was considered with regard to a single Web page. It was decided that there is no possible way to represent all forms, sizes and shapes of Web pages and Web users. Therefore there were 31 different experiments conducted in two different sessions (24 and 7 experiments, six months apart) to enhance the reliability of the results. It was also decided that the experiment should be conducted online in order to better simulate the user's interaction with online texts.

Subjects were each presented with a single Web page and a corresponding set of descriptions of the given Web page (snippets of text retrieved by InCommonSense). Each subject was asked to read the Web page and the corresponding snippets and assign them a value between 1 (bad) and 5 (good). Thus, each description was independently assigned a value between 1 and 5. The Web pages and their corresponding sets of descriptions were presented randomly (both the order of snippets on the experiment page, and between the different tested target Web pages). In order to be able to give scores with respect to the actual Web page, a different browser window was opened with the actual Web page described by the titles and descriptions annotated.

746 different subjects participated in the 31 experiments. In order to minimise the problem of not being able to talk to and direct each subject independently in an online experiment, subjects were invited to participate through strictly moderated mailing lists (such as web-research-uk, collab, CHI-WEB, diglibns, web4lib etc.), where there is an academic interest in participating in such experiments. Overall there were 252 snippets of text rated for their descriptive value.

Since the scale used in the experiment was 1 to 5, there was a need to determine how to interpret the results: For each rated snippet of text, a mean score and a confidence interval (assuming 95% confidence) were calculated. Then snippets were sorted into three groups - good examples, bad examples, and examples which are not distinctively good or bad (mixed score). The decision as to which textual entity was bad or good was based on whether the mean plus/minus the interval crossed the value 3 (on a 1 to 5 scale). If the mean stayed above the threshold, then it was considered good, while if it stayed below the 3 threshold, it was considered bad. Any result that crossed the 3 threshold was considered mixed. The decision was based on the following division:

```

if
  mean - confidence interval ≥ 3
  then good example
else if
  mean + confidence interval < 3
  then bad example
else
  in between (mixed)

```

This division (also illustrated in Figure 4) was suggested in order to certify that there is a distinct and statistically valid mechanism to identify subjects' preferences.

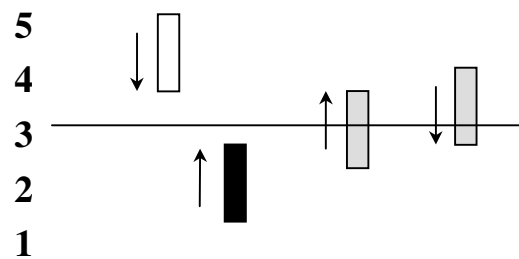


Fig. 4. Four possible situations of mean +/- confidence interval span (white = good, black = bad, gray = mixed).

4.2 Selecting features and building a filter

In order to build an automatic filter, the data from the experiment was analysed to identify language features that could be detected automatically. These features were observed in the data, ignoring the descriptive value of the texts. Initially more than 60 features were identified concerning descriptions' length, punctuation, use of verbs, position of verbs, use of adjectives, use of personal pronouns, frequency of n-grams, repetition of terms across descriptions (agreement between annotators), etc. Then, each snippet of text was analysed automatically, and a list of matching features was generated. This data was fed into a commercial, off-the-shelf, machine learning tool (See5 - corresponding to C4.5 [22]).

The goal of this stage was not to achieve the best decision about the value of a given description, but to eliminate the selection of bad descriptions, so that the system would almost never return a bad description for a search result. After optimising the features, costs, and data (190 training cases from the first set of 24 experiments and 62 test cases from the second set of 7 experiments), the classification tool never identified bad descriptions as being good. The number of features used was also reduced to 15, accounting for length, punctuation (commas, dashes, exclamation marks), use of personal pronouns, use of acronyms, use of terms expressing opinion (e.g., best, comprehensive), use of terms indicating content (e.g., about, information), position of punctuation (beginning, end), position of verbs (beginning), text beginning with capital letter, and term repetition ratio (in %).

With the aid of the See5 tool, 16 rules were hard coded in the InCommonSense system, creating a fast and independent descriptions filter. For example, description number 11 in Figure 3 was chosen by the system as best describing the URL found in <http://www.geocities.com/CapitolHill/6228/>. The system then produces the description in Figure 5. This choice was made based on the fact that this description contains some punctuation marks, begins with an 'is', and has more than 15 words. The system also found that none of the rules for detecting bad descriptions apply (e.g., words used in this description are mentioned by other descriptions, there was no excessive use of opinion terms, there was no use of personal pronouns etc.).

[Elections USA](http://www.electionsusa.com) is an independent site that strives to offer a great deal of objective information. Includes a special section on the Clinton-Giuliani Senate race in New York, along with current poll data and a big list of links.
source: <http://www.portland.com/newslinks.shtml>

Fig. 5. Best description chosen by InCommonSense for <http://www.geocities.com/CapitolHill/6228/>.

The gathering of descriptions and the filtering process is done for each URL returned by a given search engine in response to a query. Currently, InCommonSense processes 10 results at a time. Since most search engines store some information about documents in their system, it would probably be useful to store descriptions and reproduce them whenever a document matches a query. This information could be updated when the search engine crawls the Web, particularly when it is making use of popularity algorithms [12] [9].

5. PRESENTING SEARCH RESULTS WITH HUMAN AUTHORED SUMMARIES

Search engines currently use two different techniques for determining what a site is about: they either employ human editors to read the content of the site and describe it, or that they scan the site automatically and try to extract keywords and key-phrases with statistical tools. The problem of presenting this "aboutness" in the search results, in a coherent textual manner, tends to be ignored by many researchers, partly because the answer seems to be subjective and user dependent, and partly because some people assume that very "heavy" NLP techniques should be applied for achieving coherent textual summaries. As we demonstrated above, it is feasible to collect statistically significant preferences about the content of text snippets from users, and then re-use these in the context of search results.

In this section we will describe an evaluation of the InCommonSense output (i.e., short summaries of Web pages), comparing it to the output of current large-scale search engines.

5.1 Evaluation of results compared with current descriptions given by search engines

We decided to compare the summaries produced by our system with two techniques commonly used by current search engines. The first is when the top X words are taken from the document, and the second is when query terms are highlighted and the

surrounding context is taken [13]. We called the first *AltaVista Style*[29], and the second *Google Style*[30].

5.2 Method

We set ourselves a real world problem: people need to choose one result from nine possibly related results, assuming that isolating or distinguishing a single result from the others presented is not similar to a real world situation. This is because search engines are able to come up with several answers to a query, and thus InCommonSense is designed to help users in making a more informed choice. In order to compare between the three types of display (i.e., InCommonSense, AltaVista Style, and Google Style) we designed yet another online experiment to test which type of textual presentation is preferred, in terms of user satisfaction and perceived ease of interaction.

Queries. We chose short (1-3 words) coherent and general-concept queries, like names of famous people, popular leisure activities, research interests, health issues, factual information etc. The queries were partly selected from haphazard terms appearing in previous pages retrieved by our system, and partly reflect our own research and personal interests. We have tried to accommodate for a wide variety of subjects, baring in mind that most queries should be short and precise, so that sites found would relate (at least loosely) to the query. Obviously, the results we used in this experiment were different from the snippets we used to train the descriptions filter. The results we used for the final evaluation were automatically retrieved with the complete InCommonSense system.

Number of results. Starting from 10 results per page, we tried to be left with as many as possible, omitting duplicates given by the search engine (same site but a slightly different URL), broken links (404 Errors). We were left with 9 results per page.

Google Style. Highlighted terms were generally taken "as is" from the Google search engine results. When these were not available, the first two occurrences of the query terms were pasted and highlighted. Since there are various ways of displaying the short contexts around the highlighted terms, we chose to display the text on a single continuous line.

AltaVista Style. When available, results that were described in AltaVista were pasted "as is". When such snippets were not available, the first lines of text (visible - not in HTML code) were taken, up to a certain number of words.

Not showing URLs and titles. Since many people "by-pass" the lack of information by using hints from the names appearing in the URL (like cnn.com, breastfeeding.com, whalewatching.com, etc.), it was decided that such information would interfere with the goal of the experiment. As well as removing the URLs from the text, a command was added so that the bottom of the browser would not show the URL either. Titles were also removed from the results in order to make participants focus on the snippets alone.

Each participant was assigned a single task (a random task out of five). A task was composed of a short information need description and a reason for choosing the query we assigned it. After reading the task and the assigned query, a set of results was displayed with either an AltaVista Style, a Google Style or an InCommonSense (summaries) Style of display. Each participant was asked to choose the one, most appropriate, result.

After choosing a single result and viewing its content, participants were asked to answer four questions (based on and modified from [14] and [21]), rating them on a 7 point scale.

1. Are you happy with this result
2. It was easy to find the information I needed
3. I read the textual snippets given with the results in order to make a decision
4. I usually read the textual snippets given for search results.

Each interaction was recorder in a log for IP number, time, and choice, allowing us to eliminate people who participated more than once from a single IP, or that were going back and forth between screens.

5.3 Results

738 people with different IP numbers have participated in the experiment. This is after removing many duplicities, spam, and errors from the data. As far as we know, this is the largest number of participants that were recorded in published work of this type.

From our data it seems that people were happy with the result they chose regardless of the display (no significant differences). This result might be explained by the fact that participants were not able to submit their own query terms, and the task was an artificial setting. It may be that when we are able to present those same types of display on a commercial search engine, with live and real searches, that the satisfaction test will yield different results. Answering our third and fourth questions, participants also claimed to have read the results' snippets with no significant differences.

We observed that a larger proportion of the participants spent more time reading the snippets produced with InCommonSense than with the AltaVista Style or Google style results. However, this result is not statistically significant.

The most significant result we found was that for the second question - how easy it was to make a decision based on the snippets - there was a significant difference between people that interacted with InCommonSense and people that interacted with the other styles. This difference is shown in Table 1.

Table 1. Values on a 7 point scale for the question "how easy it was to find the information needed"

('ICS' stands for InCommonSense)

	mean	standard deviation	confidence	P value against ICS
AltaVista Style	4.14	1.76	0.23	P = 0 significant
Google Style	4.13	1.65	0.2	P = 0.0002 significant
ICS	4.71	1.67	0.21	

p > 0.005 for Google style - AltaVista style (not significant)

These results mean that in terms of ease of interaction, the textual output of InCommonSense is superior to the output currently used by commercial search engines.

Another interesting pattern in the data indicates differences in choice distribution between different outputs, but in order to assess the significance of such patterns, we need to design a wider collection of queries and tasks, so that the trends could be captured.

6. CONCLUSIONS

In this paper we have described a system, InCommonSense, that offers an alternative approach to summarising Web pages. This approach is based on language use conventions, practised by Web authors, in terms of layout patterns and simple language constructions. Similar notions to the ones we use here, particularly that writing conventions are useful for information extraction, are the core of the CiteSeer project [10]. However, the conventions CiteSeer follows are only the ones documented in writing style guides, and the documents it analyses are restricted to academic articles. Also, CiteSeer does not attempt to produce summaries of the referenced documents. Our system, on the other hand, aims to capture any document on the Web that adheres to the anchor-paragraph arrangement described earlier. InCommonSense does not require any knowledge about the genre of the document, nor its origin.

We are currently looking into other applications that might be improved by our approach. Such applications are editing tools for large online directories, automatic portal creation, indexing systems, and intelligent bookmark tools. We are also investigating the application of such an approach in the field of personalised information tailoring. Furthermore, in our data we found that there are many non-English descriptions for sites. Some of these descriptions are as coherent and logical, as their English parallel. In the future it would be interesting to offer these descriptions in several languages, according to the language of the user.

For more information and examples please visit: <http://www.ics.mq.edu.au/~einat/incommonsense/>

7. ACKNOWLEDGMENTS

The authors would like to thank Stephen Green (Sun Microsystems Labs), Michael Johnson (ICS, Macquarie U.), Jon Oberlander (Informatics, Edinburgh U.), and Ross Wilkinson (CSIRO) for their ideas, support, and encouragement.

8. REFERENCES

- [1] Abracos J. & Pereira-Lopes G. (1997). Statistical methods for retrieving most significant paragraphs in newspaper articles. In ACL/EACL Workshop on Intelligent Scalable Text Summarization, pages 51-57.
- [2] Amitay E. (1997). Hypertext - The importance of being different. MSc Dissertation, Centre for Cognitive Science, Edinburgh University, Scotland. Also Technical Report No. HCRC/RP-94.

- [3] Amitay E. (1999). Anchors in context. in Words on the Web - Computer Mediated Communication, Lynn Pemberton & Simon Shurville eds., Intellect Books, UK.
- [4] Amitay E. (2000). Trends, Fashions, Patterns, Norms, Conventions... and Hypertext Too. CSIRO TR 66/2000
- [5] Baldonado M.Q.W & Winograd T. (1998). Hi-cites: Dynamically-created citations with active highlighting. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Los Angeles, CA.
- [6] Barzilay R. & Elhadad M. (1997). Using lexical chains for text summarization. In ACL/EACL Workshop on Intelligent Scalable Text Summarization, pages 10-17.
- [7] Boguraev B. & Kennedy C. (1997). Saliency-based content characterisation of text documents. In Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization.
- [8] Boguraev B., Kennedy C., Bellamy R., Brawer S., Wong Y.Y., & Swartz J. (1998). Dynamic presentation of document content for rapid on-line skimming. In AAAI Spring 1998 Symposium on Intelligent Text Summarization
- [9] Brin S. and Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the 7th World Wide Web Conference (WWW7), Brisbane, Australia
- [10] Giles C.L., Bollacker K., Lawrence S. (1998). CiteSeer: An Automatic Citation Indexing System, DL'98 Digital Libraries, 3rd ACM Conference on Digital Libraries, pp. 89-98, 1998.
- [11] Hearst M.A. (1994). Using categories to provide context for full-text retrieval results. In Proceedings of the RIAO'94.
- [12] Kleinberg J. (1998). Authoritative Sources in a Hyperlinked Environment, Proceedings of the ACM-SIAM Symposium on Discrete Algorithms. Also appears as IBM Research Report RJ 10076, May 1997. <http://simon.cs.cornell.edu/home/kleinber/auth.ps>
- [13] Lesk M.E. (1989). What to Do When There's Too Much Information. in Proceedings of Hypertext '89, pp. 305-318, ACM Press.
- [14] Lewis J.R. (1995). Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. International Journal of Human-Computer Interaction 7:1:57-78
- [15] Liddy E. (1993). Development and implementation of a discourse model for newspaper texts. In Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication.
- [16] Lin C.Y. & Hovy E. (1997). Identifying topics by position. In Proceedings of the ACL Conference on Applied Natural Language Processing, pages 283-290.
- [17] Lin C.Y. (1998). Assembly of topic extraction modules in summarist. In AAAI 98 Spring Symposium on Intelligent Text Summarization, pages 53-59.
- [18] Mahesh K. (1997). Hypertext summary extraction for fast document browsing. In AAAI-97 Spring Symposium on Natural Language Processing for the World Wide Web.
- [19] Marcu D. (1997). The rhetorical parsing of natural language texts. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pages 96-103.
- [20] Paradis F. (2000). Information Extraction and Gathering for Search Engines: The Taylor Approach. in RIAO (Recherche d'Informations Assistée par Ordinateur), Paris, France.
- [21] Perlman G. (Online). Web-Based User Interface Evaluation with Questionnaires. <http://www.acm.org/~perlman/question.html>
- [22] Quinlan R. (1993). C4.5: Programs for Machine Learning. San Matco: Morgan Kaufmann.
- [23] Rau L.F., Brandow R., & Mitze K. (1993). Domain-independent summarization of news. In Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication.
- [24] Salton G., Singhal A., Buckley C., & Mitra M. (1996). Automatic text decomposition using text segments and text themes. Hypertext, 1996.
- [25] Strzalkowski T., Wand J., & Wise B. (1998). A robust practical text summarization. In AAAI 98 Spring Symposium on Intelligent Text Summarization, pages 26-33, 1998.
- [26] Tombros A., Sanderson M., & Gray P. (1998). Advantages of query biased summaries in information retrieval. In AAAI 98 Spring Symposium on Intelligent Text Summarization, pages 44-52.
- [27] Zechner K. (1996). Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In Proceedings of the International Conference on Computational Linguistics.
- [28] Zhou J. & Tanner T. (1997). Construction and visualization of key term hierarchies. In Proceedings of the ACL Conference on Applied Natural Language Processing, pages 307-311.
- [29] <http://www.altavista.com>
- [30] <http://www.google.com>
- [31] <http://www.hotbot.lycos.com>
- [32] <http://www.infoseek.go.com>
- [33] <http://www.yahoo.com>
- [34] <http://dmoz.org/>