

InCommonSense - Rethinking Web Search Results

E. Amitay

ICS, Macquarie University
NSW 2109, Australia

CSIRO/MIS
North Ryde, NSW, Australia

Abstract - The World Wide Web is a rich annotation system which allows people to relate to documents and sites from different perspectives. People describe, comment, relate or mock other Web pages in the context of their document. This richness is currently not reflected in snippets presented by Web search engines, where a search result is represented by the text found in the Web document alone. This paper proposes a new method for representing documents in Web search engines' results. This method is based on recent trends in search engine technology and provides descriptions of the retrieved documents, assembled from people's commentary and annotations on the Web. This paper suggests a new way for automatically retrieving and reusing people's annotations on the Web, incorporating these annotations into a search engine for creating a hybrid directory-search-engine, allowing for both automatic retrieval and on-the-fly human authored summaries.

I. INTRODUCTION

It is common knowledge that many users are overwhelmed by the amount of information returned by Web search engines [8]. Many of them complain that there is a need to 'visit' each result in order to find out whether or not it is relevant. This process is tiring and time consuming, which is probably why some people find using search engines a difficult task. Although commercial search engines attempt to give the document title and some textual information for each search result, there is very often insufficient information to assess the relevancy of the result without going to the document itself and reading it.

A. Presenting Web Search Results

The problem of presenting Web search results in a coherent textual manner tends to be ignored by many researchers (however, see [5]), partly because the answer seems to be subjective and user dependent, and partly because some people assume that very "heavy" natural language processing techniques should be applied for achieving coherent textual summaries. This paper suggests a simple and robust answer which relies on language usage conventions on the Web, the nature of hypertext annotations, and the fact that people have common language preferences for online textual descriptions.

There are generally two techniques for determining what a site is about: the first is to read the content of the site, which requires human effort, and the second is to scan the site automatically and try to extract keywords with statistical tools such as [6] and [10]. The first technique is

used by Web directories, like Yahoo![15], where professional people read and summarise what they have read in order to help other people find their way around a given subject. The second technique is used by Web search engines, where all documents are automatically indexed and retrieved in response to a user query.

This paper presents InCommonSense, a system that retrieves and assembles people's annotations and descriptions about a given Web site (a single, primary URL). The system then chooses the best description based on automatically detected language features, previously derived from users' preferences. This process is creating a hybrid directory-search-engine, where each search result is described by a coherent textual snippet. The system does not use a user profile and requires no user specific effort or interaction analysis. The rules are hard coded, and the retrieval process is based on simple conventional HTML structures.

II. THE InCommonSense SYSTEM

InCommonSense is a system that takes advantage of the paragraph convention found in Web hypertext [3]. It automatically extracts annotations, notes, descriptions and other scribbles that people write about other Web pages. The vision behind InCommonSense is similar to the one described by Landow [6], where he envisions a new form of linking through an information retrieval system that would automatically extract related information:

Other forms of linking will permit automatic data gathering, so that lists of relevant publications or current statements about [my document] created after I had completed [it] would automatically become available. (p. 85)

The metadata used by InCommonSense to identify such statements is implicit, and the only guide for extracting this information is the convention of online paragraph writing and the convention of positioning anchors in them.

A. 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.

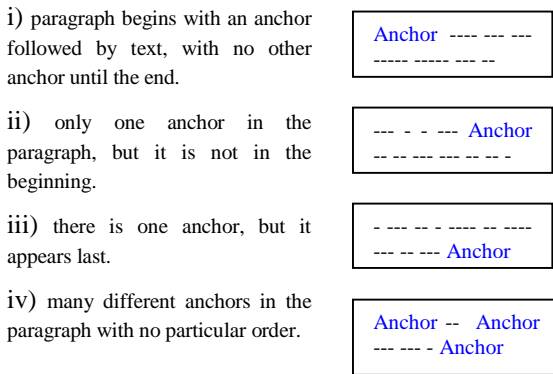


Fig. 1. Four paragraph-link patterns found on the Web.

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 [3]. Since this form of writing is practised by many users and is becoming more widespread everyday, it can be efficiently and automatically manipulated.

B. 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.

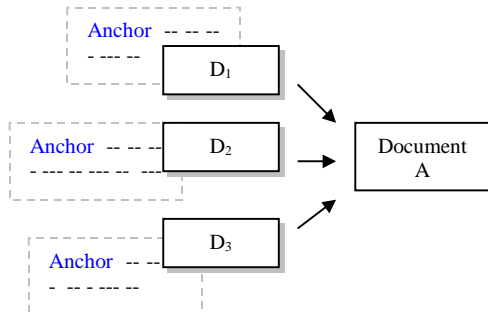


Fig. 2. Relations between documents as detected by the InCommonSense system.

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 is looking 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 [1][2]. For example, the following snippets of text (Fig. 3) relate to the document found in <http://www.westegg.com/einstein/> (titled - "Albert Einstein Online"):

(1) Anchor : includes: biography, physics, writings, quotes, pictures , and more (Source: http://www.princeton.edu/~pressman/genref.htm)
(2) Anchor : from The University of Pennsylvania (Source: http://library.aecom.yu.edu/)
(3) Anchor : - everything you ever wanted to know about one of the world's foremost physicists. (Source: http://particleadventure.org/other/othersites.html)
(4) Anchor : , in addition to having wacky hair, is a man whom I admire. (Source: http://www.westegg.com/morgan/)
(5) Anchor : -- Although we try not to review too many sites that are simply virtual libraries of information, we do occasionally come across one worth mention. Such is this Albert Einstein site--a sensibly arranged collection of links to online material about Einstein. Whether you simply want an overview, or whether you want detailed information about his physics or a short biography, this is the place to try first. And there's also a set of links to photos of Einstein, so if all you want is a copy of the famous one where he sticks out his tongue, you can get that, too! (15 October 1998) (Source: http://www.keysites.com/keysites/hotspots/physics1.html)
(6) Anchor : lists everything on the web about Albert Einstein. (Source: http://www.eaglequest.com/~bondono/astro.html)
(7) Anchor : An extensive collection of resources including links to images, texts, quotes, overviews and more Maintained by S. Morgan Friedman (Source: http://www.academicinfo.net/physics.html)
(8) Anchor : Albert Einstein: AIP History Centre (Source: http://www.aei-potsdam.mpg.de/links/relativity.html)
(9) Anchor : A really excellent collection of links to information on the Web about Albert Einstein. Pictures, writings, quotes, and general background are among the areas addressed. From here you can do everything from looking at Einstein's relativity texts to finding out how to acquire a t-shirt bearing his image. (Source: http://www.arachnid.co.uk/award/scitech.html)
(10) Anchor : provides pictures, quotes, writings and links to hundreds of other Einstein sites. (Source: http://www1.sympatico.ca/Contents/Science+Technology/science_resources.html)
(11) Anchor : - extensive collection of articles, texts and related science resources. (Source: http://www.earl.org.uk/earlweb/science.html)
(12) Anchor : - An extensive list of Einstein links, including his writings, quotes, pictures, and more. (Source: http://www.execpc.com/~shepler/einstein.html)
(13) Anchor : A site full of interesting links for the man his theory and his ideas. (Source: http://www.ultisoft.demon.co.uk/relative.html)
(14) Anchor : This site contains many links and lots of information about Albert Einstein. (Source: http://gallery.uunet.be/nicvroom/)
(15) Anchor : . S. Morgan Friedman. (Source: http://hypertextbook.com/eworld/einstein.shtml)

Fig. 3. Snippets of text that InCommonSense assumes to be related to the URL - <http://www.westegg.com/einstein/> (titled - "Albert Einstein Online").

Currently, InCommonSense processes up to 220 documents relating to a single URL in one run (using Google [12], HotBot [13], AltaVista [11] and Infoseek [14]). 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.

III. 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. 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.

A. 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.

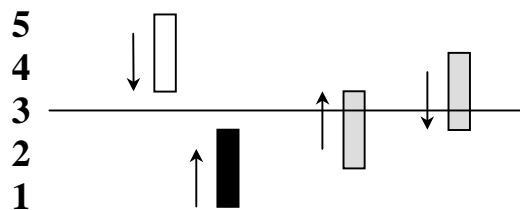


Fig. 4. four possible situations of mean +/- confidence interval span (white = good, black = bad, gray = mixed).

B. 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 [8]).

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 end), 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. The two descriptions (7) and (9) in Figure 3 were chosen by the system as best describing the URL - <http://www.westegg.com/einstein/> . From these two descriptions the system chose the shorter one, considering screen space limitations, and produces the description in Figure 5. The system decided to pick this description based on the fact that it contains more than 15 words, it expresses opinion and the text begins with a capital letter. The system also found that none of the rules for detecting bad descriptions apply.

Albert Einstein Online: An extensive collection of resources including links to images, texts, quotes, overviews and more Maintained by S. Morgan Friedman
(Source: <http://www.academicinfo.net/physics.html>)

Fig. 5. Best description chosen by InCommonSense for <http://www.westegg.com/einstein/>.

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 [4].

IV. SUMMARY AND FUTURE WORK

This paper describes a system, InCommonSense, that automatically retrieves and filters human authored descriptions of Web pages. The retrieval process is based on writing and presentation conventions on the Web, and the filtering process is based on common user preferences and deriving language based rules from these preferences. These rules are hard coded in the system but can easily be changed if, and when, conventions and preferences change. In the course of one year the system was able to use its

retrieval mechanism with no problem. Preferences were also tested over a period of six months to eliminate the temporal implications that people might change their tastes after a period of use.

For popular sites the system finds many descriptions, and it is very easy to produce a good description for such a site. Problems arise when a site is less popular, and therefore fewer people describe it. We are currently working on a method for assembling several anchors to generate a description. The process is similar, but the task is a little different since it involves assembling anchors from different sources, while the descriptions are usually written in a coherent manner by a single author.

InCommonSense might be also of use if incorporated in tools for authoring & editing portals, automatically creating directories, indexing text and non-text entities, etc. More information about the InCommonSense system and its development can be found under [16].

ACKNOWLEDGEMENTS

I would like to thank my primary supervisor, Cécile Paris from CSIRO, for her input and support of this work, and to Stephen Green from Sun Microsystems Labs for his support and ideas. I would also like to thank Michael Johnson (ICS, Macquarie U.), Jon Oberlander (Informatics, Edinburgh U.), and Ross Wilkinson (CSIRO) for their help and encouragement.

REFERENCES

- [1] E. Amitay (1997). "Hypertext - The importance of being different". MSc Dissertation, Centre for Cognitive Science, Edinburgh University, Scotland. Also TR HCRC/RP-94.
- [2] E. Amitay (1999). "Anchors in context". In *Words on the Web - Computer Mediated Communication*, Lynn Pemberton & Simon Shurville (eds.), Intellect Books, UK.
- [3] E. Amitay (2000). "Trends, Fashions, Patterns, Norms, Conventions... and Hypertext Too". CSIRO-TR 2000/66.
- [4] S. Brin & L. Page (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th World Wide Web Conference (WWW7)*, Brisbane, Australia
- [5] Hearst M.A. (1995). TileBars: visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. pp 59-66.
- [6] J. Kleinberg (1998). "Authoritative Sources in a Hyperlinked Environment". In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.
- [7] G.P. Landow (1991). *Hypertext: The Convergence of Contemporary Critical Theory and Technology*. Johns Hopkins University Press, 1991.
- [8] A. Pollock, & A. Hockley (1997). "What's Wrong with Internet Searching". *D-Lib Magazine*, March 1997.
- [9] R. Quinlan (1993). *C4.5: Programs for Machine Learning*. San Matco: Morgan Kaufmann.
- [10] K. Sparck-Jones & P. Willet (1997). "Chapter 6 - Techniques". In *Readings in Information Retrieval*. Karen Sparck Jones & Peter Willet: (eds.), Morgan Kaufmann Publishers, CA.
- [11] <http://www.altavista.com>
- [12] <http://www.google.com>
- [13] <http://www.hotbot.lycos.com>
- [14] <http://www.infoseek.go.com>
- [15] <http://www.yahoo.com>
- [16] <http://www.ics.mq.edu.au/~einat/incommonsense/>