H-0238 (H0604-019) April 29, 2006
Computer Science

IBM Research Report

On Text Around Anchors and Its Use in Web IR

Einat Amitay
IBM Research Division
Haifa Research Laboratory
Mt. Carmel 31905
Haifa, Israel

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

On text around anchors and its use in Web IR

Einat Amitay
IBM Research Lab in Haifa, Israel
einat@il.ibm.com

In this paper we report on the study of text around anchors. In particular, this paper presents the notion of a descriptive snippet. A descriptive snippet is a unit of text that appears alongside the anchor, within a visually distinct arrangement, and describes the target page the anchor links to. We show the various new ways snippets have been incorporated into retrieval systems, serving as page summaries, retrieval units, and intelligent data mining building blocks. This paper reviews the implementation and incorporation of SnipIt, a tool for extracting descriptive snippets, within the InCommonSense summarization system ([3][4]), the Google search engine([5]), IBM's Juru search engine ([6]), and IBM's WebFountain platform ([22]). We show that the extracted descriptive snippets allow an expanded, fresh and unique recount of the pages they describe, as well as provide an encapsulated coherent unit of text suitable for both retrieval and display of high quality search results.

A quick note before we start – the study presented here brings details and examples gathered over a period of nearly ten years. Some of the URLs referenced here may not be found on the server where they were originally gathered from. If the reader wishes to read and follow the original HTML files we recommend using the WayBack Machine ([21]) kindly provided by the Internet Archive. For reasons discussed later in this paper, most URLs referenced here were popular enough at the time to be crawled and stored by the Internet Archive.

1. Introduction

This paper provides a broad overview of the research effort to study text around anchors and its use in Web Information Retrieval. The paper is divided into four parts: Section 2 surveys previous studies that led to the development of the concept of SnipIt. Section 3 provides details about the design and realisation of SnipIt as an extraction mechanism for descriptive snippets. Section 4 explores the vast amount of data gathered with SnipIt through testing on Google, IBM search and mining platforms, and on the controlled TREC .GOV collection. Section 5 brings results from experiments made with SnipIt as a basic building block in various retrieval systems. The research reported in this paper is an ongoing effort lead by a single researcher but with the help of many others and has been conducted continuously since 1996.

2. On text around anchors

It has long been suggested that text around anchors can provide some indication about the content of the linked page. However, the extent to which such text should be considered for tasks such as indexing, crawling, and ranking is a source for a great debate. This section shows, through linguistic and empirical evidence, that there is a way to solve this problem without resorting to powerful and costly techniques. By identifying the area where informative text is found, we show that in 70% of the cases text around anchors is potential noise. From the rest 30% of the cases studied, only a little more than half are relevant to the linked pages.

2.1. Background

Since the creation of the first Web search engines, the problem of what portion of the plain text around the anchor should be considered for identifying the content of the target page has been an unresolved technical issue. Many solutions have been suggested but most of them were based on long tweaking and trial-and-error processes. The most (seemingly) obvious solution practised by many is to ignore the text around the anchor altogether.

Chakrabarti et al. ([10]) introduced a notion that text around anchors is descriptive of the content of the page pointed to. Assuming that the string <a href="http://www.yahoo.com"> would typically co-occur with the term 'Yahoo' (in plain text) in close proximity, they study, on a test set of over 5000 web pages drawn from the Web, the distance to the nearest occurrence of the term 'Yahoo' around all links to http://www.yahoo.com (Table 1). A distance of zero corresponds to occurrences between the tags and . A negative distance means number of bytes before the string, and a positive distances means number of bytes after the string.

Distance	-100	-75	-50	-25	0	25	50	75	100
Occurrences	1	6	11	31	880	73	112	21	7

Table 1 - Anchor text position vs. distance in bytes (taken from [10]).

These results suggest that most occurrences of the term 'Yahoo' are within 50 bytes of the href string. Qualitatively similar experiments with URLs other than yahoo.com (where the text associated with a URL is likely to be as clear-cut, i.e. where there was a specific name like CNN associated with the web document) suggested similar results.

The term looked for in the anchor was assumed to appear on the document where the link was pointing to. This assumption may be too restrictive when trying to identify a coherent description of a Web document (so, for example, people might refer to Yahoo as being "a good place to look for information"), but it still gives an indication as to the informativeness of the anchor and the free text around it. Examining Table 1 one can also see that there are more occurrences of the term Yahoo *after* the zero position than before it. This means that when we look for information around the anchor we are more likely to find related information in the text immediately after the anchor than the text immediately preceding the anchor.

Davison [13] studied the relation between terms appearing in the pointing page and the content of the target page. One of the findings in his study is that text surrounding the anchor has a much lower likelihood of being present in the target page. However, the author states from observation that there are times when text around anchors is very useful for detecting relatedness.

Both of these empirical studies are treating text around anchors as a string of potentially related terms. The relatedness of these terms to the anchor and the target page is measured by the term's distance from the anchor tags. This paper proposes a different approach that is motivated by linguistic notions (described in detail in [4] and [5]). This new approach is based on the layout of the paragraph where the anchor and the text appear, and also on the position of the anchor within that paragraph.

2.2. Linguistic notions and anchor text

Hypertext anchors are usually short noun-phrases ([1][2]). This relates to the naming phenomenon discussed by Clark et al. ([12]). To minimise the effort of conveying a clear topical message, Clark et al. propose that writers invest more thought in editing, reformulating, and refashioning the initial reference, creating a clear and simple noun phrase. Clark et al. call this process of minimised mutual effort the "*principal of distant responsibility*", where the writer of a text is responsible to convey the topic of the message in the most coherent way possible by introducing an initial accurate, clear, and simple noun phrase. Since HTML poses a physical challenge to writers of hypertext, forcing them to name the link, this naming process becomes even more visually acute. It is obvious from the literature that at least at the word level, people choose more content-indicative names as anchor text ([9][10][13][15]).

This leads us to consider anchors at the beginning of a paragraph to be titles or paragraph headings which were found to facilitated recall of unfamiliar topics ([19]). To take this even farther, in Web hypertext writing the anchor at the beginning of a paragraph sometimes replaces the topic sentence altogether. This transformation of the topic sentence into an anchor might seem only a slight one, but in fact, its visual characteristics cause people to detect and reproduce the pattern more easily. As will later be demonstrated, this pattern is so easily spotted that it can be retrieved automatically.

The use of language in Web hypertext was also found to be slightly different. Language use patterns were examined in a study of a 1000 personal Web pages ([1][2]). It was found that active verbs and simple, factual language are used in most personal home pages, and that vocabulary choice is limited to simple, unsophisticated words. A more recent study by Haas et al. ([14]) presents analogous findings. These reports suggest that the layout or shape of text is preserved from the paper form and so are the placement of the topic sentence and the single main idea. However, the topic sentence itself might be replaced by a linked anchor, the language is more simple and factual, and because of screen limitations paragraphs seem to be shorter.

2.3. Why is this arrangement of anchor and paragraph interesting?

The set of language conventions that guide paragraph writing in Web hypertext is very useful for automatic retrieval. It is useful because it puts some restrictions on the general length of paragraphs. It also assumes some content unity, i.e. there is a single topic in each paragraph. The language used in the paragraphs is simple, and the

topic of the paragraph is stated at the beginning. The most crucial issue is that since the anchor marking the link is easily identified as a naming entity, it is a good starting point for detecting the topic of the paragraph.

On the Web there are different patterns of linking within the limits of a paragraph (i.e. when the anchor appears surrounded by text). These linking patterns can generally be viewed as four distinct patterns: The first pattern is where the paragraph begins with an anchor followed by text, with no other anchor until the end. The second pattern is where there is only one anchor in the paragraph but it is not positioned at the beginning. The third pattern is where there is one anchor but it appears last, usually with anchors such as: *'click here'*, *'more information'*, *'full story'*, etc. The fourth pattern is where there are many different anchors in the paragraph with no particular order. Figure 1 shows the four different possible arrangements of anchors within a single paragraph of text. Table 2 brings examples of anchor positioning within paragraphs from real Web pages. It is proposed here that we can learn something about the content of the paragraph and the anchor by simply looking at the way the anchor is embedded and positioned in the text.

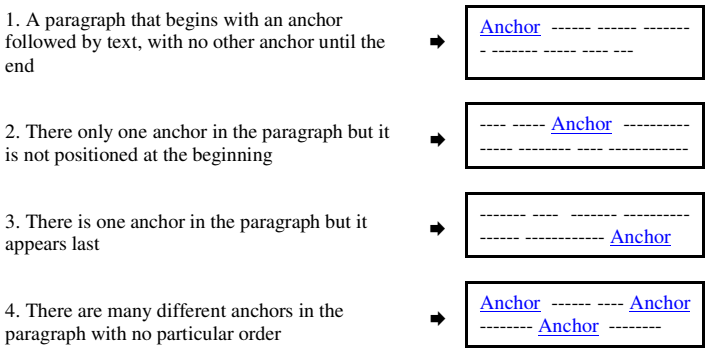


Figure 1 Four possible ways for embedding anchors in a paragraph of text.

Pattern	Example
No. 1	The Visible Human Project is supporting research to develop comprehensive, three-dimensional images of the human body. The project's web site includes cross-sectional views of the body and pointers to various medical visualizations.
No. 2	a) Did you know that an earlier version of the Labyrinth script exists? Some of the mailing list members (Walter, Stephanie, and Monica, I believe) have paper copies of it, and now Walter's scanned a text copy for your perusal. Written by Terry Jones and Laura Phillips (who is thanked in the credits), it has a lot of differences from the final version (for instance, Toby's name is Freddie). Hopefully I'll be buying a copy myself soon. b) The use of instructional technology as a classroom tool for improving upon course curriculum and instruction is an unproven tool. At present there is lots of potential, talk, workshops and ideas but few actual implementations. There is also an unfortunate tendency for potential practitioners to shy away from implementation because they are unclear if there will be an effective increase in either the quality of the instruction and/or student learning. Such apprehensions are natural and can only be assuaged by objective inspection of a few test cases. A current discipline-specific listing of Web based courses offered by various universities can be found at the The World Lecture Hall. This is an important resource for anyone considering developing Web pages as the principle lecture format. One of the main lessons to be learned is that you don't have to reinvent the wheel but instead can use and build upon what others have developed. In this case, a truly collaborative curriculum can be developed by assorted experts in the field.
No. 3	If you would like to hear The Swedish Chef explains "what happened" to Kermit click here.
No. 4	If Mark Twain were alive today, he'd probably be publishing interactive novels on the Web and charging us a fee to read them. Like many people of his time, Twain embraced new technological developments and saw them as a measure of human potential. He wrote the first novel in America to be written on a typewriter (Tom Sawyer). One of the first telephones in Hartford, Connecticut, connected the Clemens household with the central switchboard. Twain also invested (and lost) thousands of dollars in the Paige Typesetting machine, which was supplanted by the Linotype just as, nearly a century later, 8-tracks were supplanted by cassettes. Given his interest, it is no surprise to find so many references to technology new to Twain's era in his novels.

Table 2 Examples for the four different paragraph-anchor arrangements.

From studying manually tens of thousands of such examples over a period of three years (retrieved both manually and automatically) we propose that the function of the anchor in each arrangement is different. From comprehensive, long-term observations of a large volume of examples collected and stored on the local server (more than 13MB of anchor-paragraph segments from more than 250,000 different Web pages), emerge the following patterns: Almost every anchor text is relevant to the sentence it appears in, and most of the anchors are relevant to the topic of the whole paragraph.

For example, when people introduce themselves they may add links to their place of birth, where they work, what their hobbies are, etc. All of these links and anchor texts are relevant to the topic of the paragraph, which is the person described. However, the reverse relation, that the paragraph is relevant to the topic of the anchor, is not the case in this example. So, if we look at Table 2, example number 4, it is obvious that the *typewriter*, the *telephone*, and the *typesetting machine* anchors are relevant to the topic of the paragraph (Twain's fondness of technology), but the topic of the paragraph does not provide any information about the topic of these anchors and the documents they represent. The relation, in this case is unidirectional: The anchors are tools in the description of the subject, and the subject of the paragraph does not add to the description of the anchors (or the documents they represent). This case is called multi-anchor paragraph, where there is more than one anchor in the paragraph, and most of them are embedded in its middle.

Another arrangement is the one presented in Table 2, example 2a, where there is only one anchor in the paragraph and it appears somewhere in the middle. From observing many paragraphs like this one, it seems that this arrangement is very similar to the multi-anchor paragraph arrangement. It is easy to imagine that all the mailing list members mentioned in example 2a would be linked to (and their names will become an anchor). However, the case here is different, since the information can now be applied both ways. One can learn about the anchor and the link from the paragraph, and the anchor provides information about the topic. There are many facts in the paragraph that are not directly related to the anchor itself, like the names of the mailing list people, the fact that the readers should know about the first version of the script, etc.

The example of a similar arrangement in Table 2 example 2b, shows there are cases when the single anchor in the middle of the text functions exactly in the same manner as the anchors described in example number 4. The paragraph does not describe the anchor and the anchor is a tool for conveying more information and further reading about the topic of the paragraph.

From these observations (and by studying many similar cases) we can draw two general conclusions about the arrangements in examples 2 and 4: The first conclusion is that in the case where there is more than one anchor in the paragraph, it is generally not useful to take the text in this paragraph for providing information about the target page. This is so because the anchor is a reference to information displayed on a different page, or a set of pages. The second conclusion is that when there is only a single anchor in the middle of the paragraph, it is not always the case that the paragraph contributes information about the document the anchor represents. It may or may not be related to the topic of the text, and it may or may not be useful to take the text to be indicative of the content of the linked page. This ambiguity exists for all the examples studied where there is only one anchor in the middle of the paragraph.

The other two arrangements that exist are almost interchangeable. The first pattern (Figure 1, pattern 1) and the third pattern (Figure 1, pattern 3) are informative both ways: the topic of the paragraph is the topic of anchor, and maybe even the topic of the page or site that the anchor represents. Therefore the text in the paragraph is useful in the context of the page it appears on and in the context of the target page.

Although both first position and last position paragraph-anchor arrangements are useful, they do not share the same linguistic characteristics. From the data collected it appears that when the anchor is positioned at the beginning of the paragraph, the structure of the message is simple and fluent. However, when the single anchor is placed at the end of the paragraph it is usually more deictic (*click here*, *more information*, etc.), and less informative as a title. Also, it seems that this pattern is rarer and practised less and less when we examine the data collected over time.

3. Automatically extracting descriptions from paragraph-anchor arrangements: a proof of concept

In order to examine the usefulness and the frequency of each of the paragraph-anchor arrangements we implemented an extraction mechanism for retrieving such arrangements. This mechanism, named SnipIt, retrieves all of the four different arrangements. The extraction process is completely automatic, and is guided by simple markup that form the layout of HTML marked-up documents (Web pages).

3.1. SnipIt - the system

SnipIt is a simple extraction system that looks for a given URL in an HTML marked-up page, identifies its position in a paragraph of text, and retrieves the anchor naming the URL together with the paragraph of text that appears in close proximity. In order to be able to use the extracted information, and to couple the extracted text with a certain entity or concept, each URL is considered a potential entity (e.g. a Web page, an audio file, an image, etc.), that is described by the retrieved text pointing to it. The idea is simple: Since many people refer, in the context of their own page, to other Web pages, we have a mechanism that extracts information about a given Web page from any page that points to it.

For the purpose of explaining SnipIt and its implementation let us define several concepts and terms which would be used in describing the system. A *Target page* is a single HTML node (document) with a unique URL. A target can be any page on the Web. A *Pointer page* is a single HTML node with a unique URL which contains a link to the target page. From the target pages SnipIt collects the title. From pointer pages the system collects the title, the anchor which contains the URL pointing at the target page, and the paragraph of text this particular anchor appears in. An example of a target page and the information collected with SnipIt from its pointers can be viewed in Table 3.

Target page : http://www.almac.co.uk/chef/chef/chef.html Title: Ze sveedish chef	
Pointer URL	**Anchor and text**
people.ne.mediaone.net/amalkin/links.html	Swedish Chef Page: Has an encheferizer which converts your text into mock swedish. (6)
www.ibiblio.org/dls/raindex.html	Eek! Check out a page all about Ze Sveedish Chef! Bork bork bork! Or read all the dumb celebrity gossip at The Datalounge. Wanna see a picture of myself and my sister Jill before we got put into the institution? Don't look too closely, because I look like hell in that photo.
www.netsurf.com/nsd/v01/nsd.95.04.03.html	Usenet feteruns vill unduoobtedly hefe-a nuticed un oobscoore-a noosgruoop celled alt.swedish.chef.bork.bork.bork. Zee Svedish Cheff pege-a brings zee primery fucoos ooff zee gruoop tu zee VVV - a prugrem celled zee Inchefferizer. It is used tu toorn oordinery text (like-a this erticle-a) intu zee lungooege-a spukee by zee belufed Svedish cheff frum zee Mooppet Shoo. Zee pege-a elsu cunteins cheff pics und suoonds (bork, bork, bork!) und a FAQ ebuoot zee prugrem, vhich is efeileble-a fur muny pletffurms. "http://www.stir.ac.uk/~sac06/chef/chef.html"
www.people.fas.harvard.edu/~olmsted/kcole/swedish.htm	and no Swedish page would be complete without a mention of the Swedish Chef - Ze sveedish chef page, Bork Bork Bork.
www.ph.tn.tudelft.nl/~dick/	The Swedish Chef re-mailer is based on software written by John Hagerman, who maintains the Chef's homepage.
www.smartlink.net/~hiller/pghumor.htm	Ze sveedish chef, Bork Bork Bork - Home page for the Swedish Chef from the Muppets show with image and sound files and an interactive encheferizer for converting text into Chef speak
www.dailyglobe.com/jrb4.html	Swedish Chef-English Convertor Yes, the one from the Muppets
pegasus.cc.ucf.edu/~cjr71244/links.html	Ze sveedish chef - The Hilarious chef from the muppets show.
www.nationalpost.com/content/features/webhound/food.html	Ze sveedish chef page Remember ze Sveedish chef from the Muppets? ("Put ze cheeken in de oven. Bork de bork, de bork de bork bork bork.") Here you can reminisce with pictures and sounds of the chef. There is a usenet group for trading chicken recipes and talking in mock Sveedish. Or, you can download the "Encheferizer," a linguistic translator with programs available for various software: Dee en-chef-erizer mekes noormal teext look leke dee Swedish Chef teext. Bork Bork Bork!

Table 3 pointers collected for "*Ze sveedish chef*" target page

3.2. Tool design

SnipIt was designed to collect data from pointers and targets for the purpose of generating a list of extracted items for each page as described in Table 3. In order to achieve this, SnipIt accepts as input a URL (inserted by the user in the command line). Then SnipIt sends a command to a search engine querying for links to this target URL. Querying for pointers to a URL (sometimes described as "incoming links") is a feature that exists in most commercial search engines. Its syntax is generally similar to `link:URL` (where the URL is replaced by any www.X.com, or a similar Web address).

The search for paragraph text around the anchors is performed within *blocks*. A block is defined as text delimited by HTML new-line commands such as <p>,
, <li>, etc. Other commands which imply new-line are taken into account too, e.g. <h1>, <hr>, <table>, etc. The system looks for combinations of markup commands that create a visible space between the paragraphs, creating a clear-cut paragraph break on the screen (so that some markup combinations allow an anchor to appear on-top of the paragraph, but still within the visual paragraph boundary). Figure 2 describes some paragraph marks as they are hard coded into the system.

```
Text begins with <p> ends with </p> <p> <li> <hr> <ul> <dd> <dt> <img <h1,2,3,4,5,6>
</blockquote> </table> </td> </tr> (this arrangement allows for <br> inside the paragraph)

Text begins with <li> ends with </li> <li> </ul> <h1,2,3,4,5,6> <hr> </table> </td> </tr> <p>

Text begins with <dd> ends with <p> <li> <hr> <ul> <dd> <dt> <img <h1,2,3,4,5,6> </table>
</td> </tr> </th> </blockquote>

Text begins with <dt> ends with <p> <li> <hr> <ul> <dd> <dt> <img <h1,2,3,4,5,6> </table>
</td> </tr> </th> </blockquote>

Text begins with <br> ends with <p> <li> <hr> <ul> <dd> <dt> <img <h1,2,3,4,5,6> </table>
</td> </tr> </th> </blockquote>
```

Figure 2 Examples for paragraph markers SnipIt recognises

3.3. Comparing the frequency of the different paragraph arrangements

SnipIt allows for a comparison between the frequencies of the four possible paragraph-anchor arrangements. Table 4 and Table 5 present a comparison of arrangement frequencies. In order to collect a large number of examples for this comparison, SnipIt was fed with URLs gathered by querying a search engine with 25 different queries. The queries were arbitrarily chosen, and the only guide for choosing them was that they should be a single coherent concept. Each of the queries returned 10 URLs of target pages (comprising together a collection of 250 target pages). Each of the target pages was processed with SnipIt for gathering pointer pages to it, and for collecting corresponding paragraph-anchor arrangement from those pointers.

Overall 11,938 pages were collected in which a link appears to a given target page. Some of these pages contained a single anchor without any text in its proximity. In fact, only 30.7% of the pointing anchors found were embedded within a paragraph of text. All the rest of the anchors, 69.3%, appeared in isolation, without any related text around them. This finding accounts for the indecision about the usefulness of text around anchors reported in [13]. This reported uncertainty is due to the fact that the text around anchors was taken regardless of paragraph boundaries, which most probably resulted in about 70% noise.

term queried	no. of anchors	arr. 1	arr. 4	arr. 2+3	imaged anchors
albert einstein	552	85	84	8	3
antarctica	231	70	29	5	1
audio books	237	37	18	3	6
breast cancer	579	133	51	4	5
celtic	497	56	38	6	92
children books	436	112	44	3	3
currency converter	660	94	79	24	8
dictionary	457	92	45	6	2
digital library	547	120	71	18	1
edinburgh	528	56	60	4	2
elections	450	119	35	3	1
food	456	84	34	3	3
free images	402	45	71	35	5
genome project	541	84	53	7	1
hypertext	695	84	168	8	1
kurt vonnegut	489	53	29	3	1
maritime museum	470	94	28	11	1
mark twain	403	76	40	4	1
monty python	577	81	52	6	5
nlp	377	79	41	14	3
ny city guide	433	64	96	4	0
origami	465	106	55	18	1
web	498	40	49	14	13
wildlife on the web	402	62	46	5	4
world time zones	556	86	107	9	6

Table 4 frequency of the four different paragraph-anchor arrangements in 11,938 arbitrary snippets

As shown in Table 4 and Table 5, the number of paragraph-anchor arrangements of the type containing one anchor but not at the beginning of the paragraph is quite small (only 6% of all the anchors that appeared in a paragraph of text). Out of the anchors that were embedded within a paragraph of text 39% appeared with other anchors in the same paragraph, ranging from 2 to 24 anchors in a single paragraph of text. 55% of all the snippets were an anchor that appears at the beginning of the paragraph, followed by text with no other anchors. Table 5 and Figure 3 summarise the comparison between the frequencies.

Summary of results	
target pages	250
pointer pages	11,938
anchors within a paragraph of text	3660
arrangement 1 Anchor ---- --- --- --- --- --- --- -- - --- --- --- ----- ---	2012 (55% of anchors within text)
arrangements 2 + 3 ----- ----- Anchor ---- ---- --- --- --- --- --- --- --- --- and ----- ------ ----- ----- ----- ------ -- --- ------ ---- --- Anchor	225 (6% of anchors within text)
arrangement 4 ------ Anchor ----- Anchor --- Anchor --- ----- ----	1423 (39% of anchors within text)

Table 5 Percentage comparison between the four different paragraph-anchor arrangements

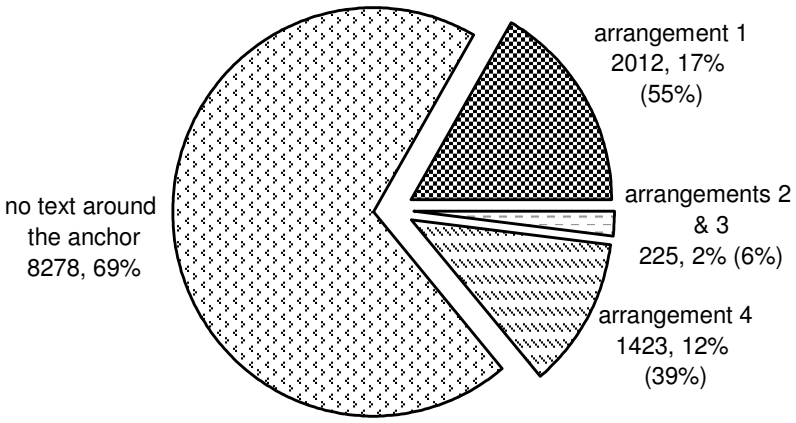


Figure 3 Frequency of anchor arrangements within a paragraph of text (numbers in brackets provide the relative frequency of anchors found with text around them)

3.4. Modifying SnipIt

From this frequency comparison it can be learnt that arrangement number 1 in Table 5 is more prevalent than the other anchor-paragraph arrangements. It can also be determined that both arrangement 2 and arrangement 3 are less likely to occur in pointer pages. Arrangement number 4 is of less use since it is hard to compute just from the layout which of the anchors is more relevant to the topic of the paragraph than the others. It was therefore decided that only arrangement 1 will be retrieved by SnipIt. This decision is based on the theoretical evidence brought earlier about the importance of the position of the topic in the paragraph.

Thus, every descriptive snippet displayed by SnipIt is a paragraph of text that appears in the pointer page. Such paragraph contains only a single anchor that appears at the very beginning of the text. This paragraph is then assumed to be a description of the anchor, and the page it points to.

3.5. SnipIt's current output

The output from the system looks like a list of pointers and their details, preceded by information extracted from the target. Figure 4 is a portion of an output file processed for the target page found under the URL: www.law.pace.edu/env/energy/globalwarming.html. The information collected is not limited to any language and many of the snippets are in languages such as German, French, Russian, Japanese, Hebrew, Spanish, etc (In Figure 4 there is one snippet extracted from a page written in Norwegian). The only guide for collecting this information is the layout markup (visual paragraph boundaries), and anchor position in the text.

TARGET:
URL: http://www.law.pace.edu/env/energy/globalwarming.html
TITLE: Global Warming Central a project of the Pace University School of Law Energy Project

---

Pointer:
http://www.dmoz.org/Science/Environment/Global_Change/
ANCHOR: Pace University "Global Warming Central"
SNIPPET: - Created by Pace University School of Law (New York) to provide "key" information for understanding the global warming debate.

Pointer:
http://www.law.pace.edu/env/energy/energyhomeproto.html
ANCHOR: GLOBAL WARMING CENTRAL
SNIPPET: - an on-line resource providing access to original information on both sides of the Global Warming Debate. The site provides key documents such as the Kyoto Protcol as well as access to global warming related speeches, experts, news and links.

Pointer:
http://www.academicinfo.net/environstwarming.html
ANCHOR: Global Warming Central
SNIPPET: Created by Pace Energy Project, a project of the Pace University School of Law, to provide 'key' information for understanding the global warming debate. Provides 'key' reports, documents, treaties, experts, and much more Maintained by Melanie Fund Pien

---

Pointer:
http://www.law.ecel.uwa.edu.au/intlaw/environment.htm
ANCHOR: Global Warming Central (Pace Energy Project)
SNIPPET: Recent News; Reports; Legislative Action; Experts; The Debate; Websites; Government Documents; NGO Documents; Treaties

---

Pointer: http://www.thimun.org/research/
ANCHOR: Global Warming Central
SNIPPET: at Pace University School of Law

Pointer:
http://www.vg.no/nyheter/utenriks/97/11/html/kyotofront.html
ANCHOR: Global Warming Central
SNIPPET: er et nettsted laget av studenter ved Page University School of law. Her finner du den informasjonen du trenger for åkunne henge med i klimadebatten (på engelsk).

Pointer:
http://www.zdnet.co.uk/athome/yahoo/1999/01/11_01_99.html
ANCHOR:www.law.pace.edu/env/energy/globalwarming.html
SNIPPET: And why is it so wet? Global warming mate. Separate the facts from the fiction at this comprehensive site.

Pointer:
http://usinfo.state.gov/journals/itgic/0498/ijge/gj-16.htm
ANCHOR: Global Warming Central
SNIPPET: Pace University School of Law provides key documents on climate change http://www.law.pace.edu/env/energy/globalwarming.html

---

Pointer:
http://www.exploratorium.edu/learning_studio/cool/environmental.html
ANCHOR:Global Warming Central
SNIPPET: - This site has everything you want to know about the growing debate over global warming, including the text of the Kyoto Protocols, news of recent scientific and political developments, and expert testimony from both sides of the debate. Cool Site: March 1998

Figure 4 example output retrieved with SnipIt for the target found under the URL http://www.law.pace.edu/env/energy/globalwarming.html

4. Experimenting with SnipIt

In this section we report on extended experiments performed on various large collections, analysing millions of pages using SnipIt. Those experiments later led to incorporating SnipIt as an integral component in the overall retrieval system. The next section will report on the uses of SnipIt in such large scale retrieval systems.

4.1. Experiments conducted by Google

Instructions and code for implementing SnipIt were given to the Google search engine research staff in 2000, in order to have it tested on the Google collection. According to their tests with approximately 400,000,000 Web documents, about 5%, of all Web pages are pointed to by descriptive paragraphs of the kind outlined here (arrangement 1).

The above estimate is based on the following findings: Approximately 400,000,000 documents were collected and examined with a SnipIt component. SnipIt found 47,114,616 descriptive snippets, out of which 32,099,877 snippets were found to have the target URL within the collection. Overall the number of targets described that were within the collection (i.e. each target may be described by more than a single snippet) was 6,975,748. The average target within the collection was therefore about 4.6 descriptive snippets per target.

We consider this a very good coverage. Compared with the current coverage of the largest available Web directory of annotated links (the Open Directory Project, dmoz.org) the coverage of SnipIt is several orders of magnitude larger. The published number of annotated URLs in the Open Directory in 2001 (when Google

conducted the experiments) was a little over 2,000,000, which means that it described less than 0.2% of the pages in the Google collection. Over 30,000 editors were required in 2001 to manually annotate and maintain this directory of links. The number of people required to manually annotate the links in the case of SnipIt was and is zero, since the system reuses annotations rather than making people annotate links especially for its descriptive purpose.

The constant growth of Google's collection and the constant growth of the Web as it is reflected in Google, reveals millions of new snippets every day. Since the maintenance of those snippets is a distributed effort performed by millions of Web authors daily, with no relation or intent of being published in a Web directory such as the Open Directory, this process of growth creates an endless collection of fresh and updated snippets. The advent of Wiki's like Wikipedia also promotes the descriptive nature of the texts authored and hence provides additional resource for snippets.

4.1.1. Correlation between PageRank and descriptive texts following anchors

Experience shows that for most queries submitted to Google roughly 80% of the top ten results will have at least one snippet describing it. The easiest way to test this is by using the 2002 live Google experiment [23]. Although the age of the collected snippets is starting to show, the top results are usually still the most prominent in terms of reputation (i.e. PageRank) and in terms of descriptive interest (i.e. snippets per page). For example, Table 6, Table 7, and Table 8 show the number of descriptive snippets ("WebQuotes" as they were named by Google[1]) per URL found by Google for the queries "*news*", "*google*", and "*star trek*".

Query: "news"			
Rank	Site	URL	Number of WebQuotes
1	CNN.com	http://www.cnn.com/	503
2	BBC NEWS	http://news.bbc.co.uk/	147
3	FOXNews.com	http://www.foxnews.com/	58
4	ABC News	http://abcnews.go.com/	31
5	Google News	http://news.google.com/	15
6	usnews.com	http://www.usnews.com/usnews/home.htm	31
7	usnews.com	http://www.usnews.com/	72
8	Wired News	http://www.wired.com/	131
9	CNET News.com	http://news.cnet.com/	12
10	CBSNews.com	http://www.cbsnews.com/	12

Table 6 Number of descriptive snippets ("WebQuotes") per URL found by Google for the query "news"

Query: "google"			
Rank	Site	URL	Number of WebQuotes
1	Google	http://www.google.com/	1581
2	Google News	http://news.google.com/	15
3	Google Groups	http://groups.google.com/	127
4	Google Weblog	http://google.blogspace.com/	3
5	Google Toolbar	http://toolbar.google.com/	34
6	Google Image Search	http://images.google.com/	51
7	Google UK	http://www.google.co.uk/	25
8	Google Canada	http://www.google.ca/	12
9	Searching Stanford	http://www.stanford.edu/search/	0
10	Google Japan	http://www.google.co.jp/	44

Table 7 Number of descriptive snippets ("WebQuotes") per URL found by Google for the query "google"

[1] WebQuotes has recently been replaced by the "define:" feature. To examine this Google feature submit a query preceded by the "define:" command. For example "define:google".

Query: "star trek"			
Rank	Site	URL	Number of WebQuotes
1	STARTREK.COM	http://www.startrek.com/	118
2	Star Trek: The Next Generation	http://www.ugcs.caltech.edu/st-tng/	7
3	Star Trek: WWW	http://www.stwww.com/	10
4	Paramount Pictures	http://www.paramount.com/	29
5	TrekToday	http://www.trektoday.com/	36
6	Star Trek: The Experience	http://www.startrekexp.com/	8
7	The STArchive	http://www.cs.umanitoba.ca/~djc/startrek/	2
8	Star Trek Nexus	http://members.aol.com/treknexus/	2
9	Bevis and Duncan's Star Trek	http://www.ee.surrey.ac.uk/Contrib/SciFi/StarTrek/STTNG/	1
10	IMDB: Star Trek: Nemesis	http://www.imdb.com/Title?0253754	0

Table 8 Number of descriptive snippets ("WebQuotes") per URL found by Google for the query "star trek"

Currently, we cannot model the exact correlation between the PageRank value of a page [9] and the number of descriptive snippets we collect that point to it. However, as a rough rule of thumb, we expect to find the top results of common queries about publicly well known concepts to have an above average number of descriptive snippets pointing to them. Through experimentation we also found that computer-related concepts tend to be described more often by Web authors.

4.2. Experiments conducted with SnipIt on the TREC .GOV Web data

So far, most of the implementations of SnipIt were used for gathering information for direct display to the user, either as a third-party summary of the search results or as a tool for exposing people to what is being said about them or their site. In 2003 we decided to test the power of SnipIt for gathering and adding meta-information to documents indexed in a retrieval system. To this end we used TREC's .GOV collection [8] which comprises 1.2 million textual documents and is 18.1GB in size. We chose to use this collection for its wide and comprehensive interconnectivity [20] which, although smaller than the 2001-2002 Google collection of WebQuotes, was much easier for processing, indexing and testing.

We analysed the .GOV collection and found 732,741 descriptive snippets. We were able to map 101,670 of those snippets to 45,406 single .GOV target pages within the TREC collection. There were 556,730 snippets describing target URLs within the .GOV domain that were not found in the TREC collection. We also found 74341 snippets describing external sites that were also outside the scope of the TREC collection. Compared with the 400,000,000 pages Google collection previously studied, the .GOV collection is very "self-centred" and only a small portion of the pages describe pages outside the domain. However, as pointed out by Soboroff [20], the .GOV collection seems to have a small "centre of attention" from which it was gathered, and the rest of the pages seem to be pointing out to uncrawled areas of the domain. Figure 5 juxtaposes the in-collection/out-of-collection ratios of descriptive snippets found in both collections. The ratio of Google's collection seems to suggest that the crawl was more comprehensive, encompassing many of the target URLs. Assuming Google's sample collection represents the actual collection of indexed pages, the descriptive snippets provide an informative glimpse into the roughly 30% uncrawled areas of the Web.

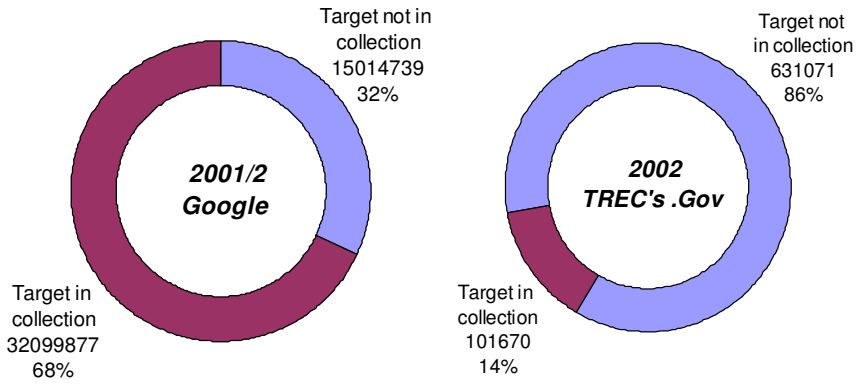


Figure 5 Percentage of descriptive snippets within the analysed Google & TREC's .GOV collections

We analysed the descriptive snippets found within the .GOV collection to provide some insight into what types of files people choose to describe. We were able to identify the type about 85% of the files described (623888 out of a total of 732741 descriptive snippets). Since only 14% of the targets were in the collection we decided to examine the single snippet and not to try and map the whole linking graph of the .GOV domain. This maybe a weakens the findings, but still provides a good indication about the preference of individual authors and their inclination to describe information in multimedia environment. The findings of this analysis are found in Figure 6. As expected, most of the files described are HTML. However, we have found many types of files to be described that can never be automatically described by text in any other means, such files included audio and video detailed descriptions, images, zip files and executable programs.

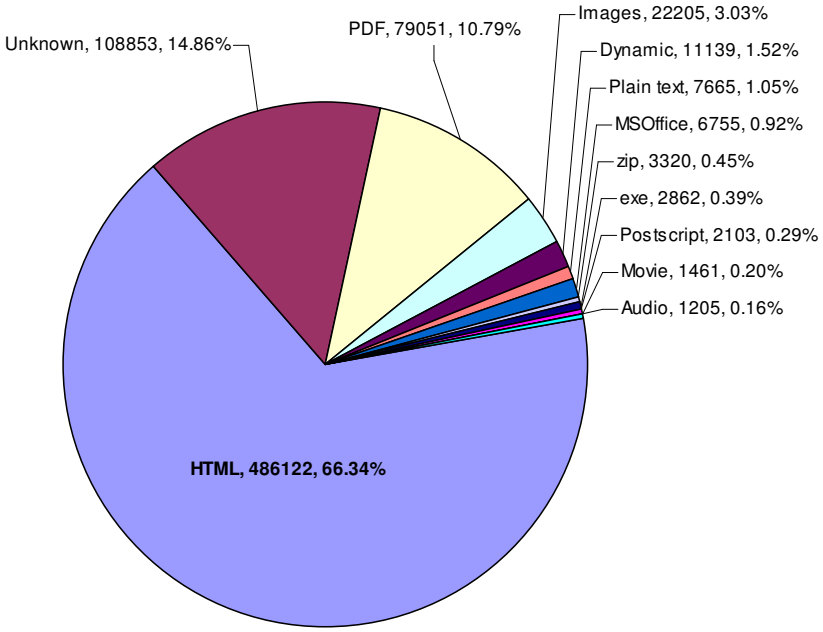


Figure 6 Analysis of the distribution of file types people choose to describe with snippets in the .GOV collection

Table 9 provides a glimpse into the thousands of examples we extracted from the .GOV collection of descriptive snippets that provide a detailed narrative describing the content of non-textually-indexable content of images, movies, audio, and executables found in the space of the .GOV domain.

Source DOCNO	Target URL	Snippet found	File type
G40-81-0950079	Tools/PLUGLOAD.XLS	: calculates annual electricity use and potential savings	MS Excel
G40-86-0732945	starting/L2_in_IVHM_training.ppt	(Power Point 2053k) Livingstone2 fundamentals, usage, and embedding in Integrated Vehicle Health Management (IVHM) systems.	Power Point
G02-75-2686174	/products/analysis_monitoring/bulletin/fige6.gif	Time series of pattern amplitudes for the last few years for selected teleconnection patterns, from the Climate Diagnostics Bulletin.	GIF
G02-75-2686174	ftp://ftp.ncep.noaa.gov/pub/cpc/wd52dg/data/indices/tele_index.nh	Monthly Tabulated Indices for all teleconnection pattern amplitudes dating back to 1950.	?
G01-44-4130921	cruisetrack7.png	Figure 3: Proposed GasEx 2001 cruise tracks. Click on above for larger image: 201,383 kb. (1020x610 vs. above 700x420)	PNG
G41-90-3785610	graphics/20001226.jpg	Craters and Boulders on Eros (Click image to enlarge)	JPEG
G01-58-3435450	http://www.broadcast.noaa.gov/ogp/ogppromo3.ram	premiered in 1998 at the Explorer's Club Annual Meeting held at the Waldorf-Astoria Hotel in New York as part of the Explorer's Award given to Dr. Ants Leetmaa, Director of the NOAA Climate Prediction Center.	RAM
G16-26-1941177	http://videocast.nih.gov/ram/mfp091900.ram	Dr. Allan Kirk Chief, Transplantation Section National Institute of Diabetes and Digestive and Kidney Diseases	RAM
G38-56-0991565	http://mam.er.usgs.gov/media/amphibians/amphibians.rm	A RealMedia File showing different amphibians. The segments show different frog species in an artificial habitat. Another segment shows the collection of salamanders in the field.	RM
G41-52-0331389	http://lunar.arc.nasa.gov/education/models/images/model-simple.jpeg	Instructions to build a simple model of the Lunar Prospector spacecraft.	JPEG
G04-32-1718895	http://vulcan.wr.usgs.gov/Imgs/Jpg/MSH/SlideSet/4.jpg	A "bulge" developed on the north side of Mount St. Helens as magma pushed up within the peak. Angle and slope-distance measurements to the bulge indicated it was growing at a rate of up to five feet (1.5 meters) per day. By May 17, part of the volcano's north side had been pushed upwards and outwards over 450 feet (135 meters). The view is from the northeast. -- USGS Photo by Peter W. Lipman, April 27, 1980	JPEG
G04-32-1675592	../../file/demoaa/sonomnd1.xls	: Total Population by Race/Ethnicity and Sex (Excel)	MS Excel
G01-36-3790183	gmedia/theme_song.ram	Listen to the Green.Gov National Theme Song (Generously contributed to the Green.Gov vision by composer Alan Menken and lyricist Glenn Slater.)	RAM
G10-59-2173649	software/znewcon.exe	is a program to assist in the design of flexible pavement structural sections for NEW CONstruction based on Chapter 600 of the Caltrans 1990 Highway Design Manual. This file is 64 Kb, Release April 30, 1991 .	EXE

Table 9 Examples for descriptive snippets pointing to file types that are otherwise difficult to textually describe.

5. Using SnipIt in Extraction and Retrieval systems

The obvious conclusion from our exhaustive analysis is that the descriptive snippets extracted with SnipIt can, and should be used in a link-based retrieval system. The reason for believing that they add information that is not present in either the linking anchors or the text of the target itself are threefold: First, some of the file types described contain no text whatsoever, and any additional text will be adequate for indexing. Second, the information provided by sources outside target text itself may provide an objective view of the page with illustration of its relatedness to the linking entity, which may be described in a contextually different environment. Third, the brevity with which people describe other pages often causes them to select only keywords that are essential for conveying the importance and relevance of the linked page to their context. This characteristic of descriptive snippets makes them good candidates for summarisation of search results and for a longer, more verbose, form of anchor text.

5.1. Using SnipIt as meta-information about the target page

The obvious use for descriptive snippets is displaying them as meta-information about the target page. Snippets may be used as a "what they say about my site" feature which show all the found description in a single list (a-la Google WebQuotes [23]). Snippets may also be used for search results display where a filter chooses only the single most appropriate snippet to be displayed.

5.1.1. Providing summaries for search results

We designed an application that makes use of the extracted snippets. This application was named InCommonSense and it is a mechanism for producing short coherent snippets to describe Web search results. InCommonSense chooses the best description out of the snippets found by SnipIt. InCommonSense is based on experiments conducted with 746 users who rated descriptions for their quality. Their preferences and choices were then trained and tested through a machine learning process. The rules derived from this process were hard coded into InCommonSense. Figure 7 shows the display of InCommonSense for the query "*Albert Einstein*" the ranking was directly taken from Google, while the display was generated by InCommonSense..

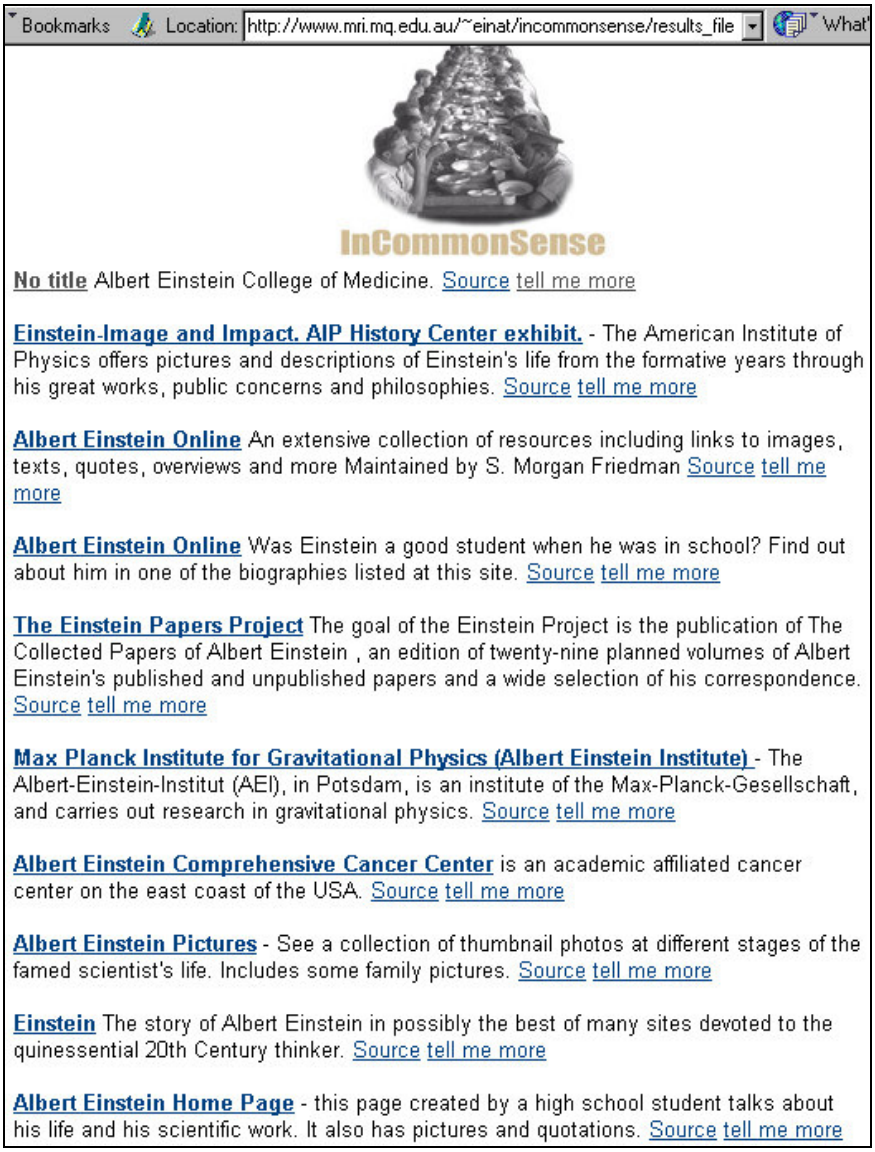


Figure 7 Search results snippets chosen by InCommonSense for the query "*Albert Einstein*". The ranking and search results were generated by Google.

The output of InCommonSense, descriptions for search engine results, was rigorously evaluated against the current output of commercial search engines. This was done by way of an online experiment with over 1000 participants. Each of the three displays we compared corresponds to a display "style". A Google style result display is the query biased summary style of display, where the query term and its context is extracted from the document found, and the query terms are highlighted in it. 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:

> ...referring to **time zones** Related links The Personal **World**....Australasia] [Personal **World** Clock][Sort by Country] [Sort by **Time**...

AltaVista style results display refers to the way some commercial search engines choose to display the first lines of the document, rounded to the word (i.e., counting the number of characters that occupy a single line on the screen, and then ending the line when a word ends rather than chopping it in the middle). Each of the three styles was presented to the user *without* any other textual hints such as the title or the URL of the page.

	mean	standard deviation	confidence	P value against ICS
AltaVista	4.14	1.76	0.23	P < 0.005 significant
Google	4.13	1.65	0.2	P < 0.005 significant
ICS	4.71	1.67	0.21	
p > 0.005 (P=0.4742) for Google style against AltaVista style (not significant)				

Table 10 Values on a 7 point scale for the question "how easy it was to find the information needed" answered by over 1000 experiment participants ('ICS' stands for InCommonSense)

Our most significant finding was the reported difference between the displays for the question: how easy it was to make a decision based on the snippets. There was a statistically significant difference between the answers from people that interacted with InCommonSense and the answers from people that interacted with the other styles. This difference is shown in Table 10 and in Figure 8. People using InCommonSense snippets said that it was easier for them to choose a result than people interacting with the other two displays. The information InCommonSense provided facilitated the choice participants had to make between the results. In terms of ease of interaction, our evaluation showed that InCommonSense is superior to the output of the commercial search engines tested. The experiments for realising the rule-based filter are described in [3]. The user experiments are extensively discussed and analysed in [5].

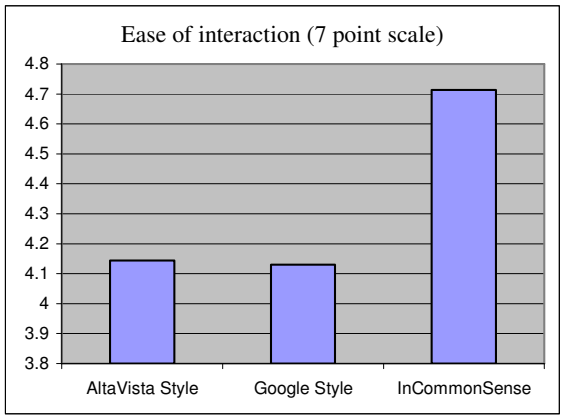


Figure 8 Statistically significant differences found between InCommonSense and the Google & AltaVista styles for question 2: how easy it was to find the information needed on the results page

Following the work on InCommonSense, there have been many systems using SnipIt-like applications to extract summaries of Web sites [14][25].

5.1.2. providing word-of-mouth for target pages on IBM's WebFountain

Similar to the way SnipIt was used in InCommonSense for summarization, IBM's WebFountain [22] allows its users to query its collection for word-of-mouth pieces. WebFountain is a platform for data mining, where each set of miners runs on the collection of pages and tags them with keys describing some meta-information related to the content of the page. SnipIt is currently one of those miners, providing a bi-directional way for querying pages: (a) what the current page is saying about all its target pages, and (b) what is being said about a target page by every page that points to it. The miner can be used for both "gossip extraction" about pages, and for spotting potential good sources for information (by using those sources that provide good description of prominent/interesting target pages).

5.2. Using descriptive snippets as retrieval units

The intuition behind using descriptive snippets as a stand-alone retrieval unit was the main reason for starting the pursuit after the exact boundaries of text around anchors and its contribution to Web IR. However, since after finding the exact boundaries we realised there is much more to this text then previously thought. So much so, that we neglected to use it for its original purpose. To come full circle we decided to participate in TREC 2003 with descriptive snippets as independent retrieval units.

We decided to run Juru ([6]) at TREC with SnipIt's output as one of its basic retrieval units. Juru has a modular built-in mechanism that allows its users to tweak the ranking score function based on the type of text it indexed. This feature allowed us to introduce the new retrieval unit into the score function for the purpose of evaluation.

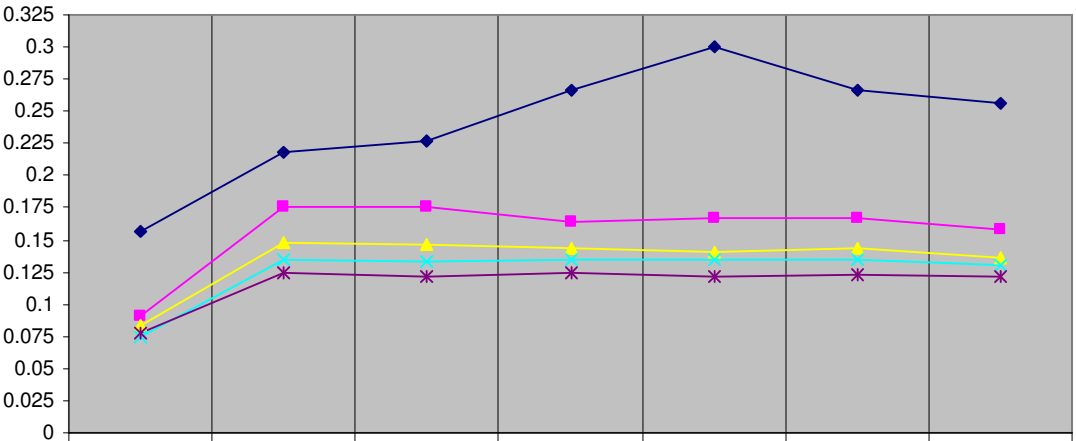
5.2.1. The TREC experiment

We used the combined set of 100 queries from the topic distillation tracks of TREC 2002 & 2003. We merged the two sets of test queries based on the assumption that the change in results may be more convincing and more prominent. We also used the .GOV collection analysed in section 4.2 above. As stated earlier, this collection seems to be biased towards a central seed of well connected pages and with many out links that were not crawled. This characteristic of the collection manifested itself in the number of descriptive snippets that were indexed and the number of target pages described. For evaluation purposes and since only those pages that were crawled were used in TREC's evaluation pool, we added to our index those snippets that described documents within the collection (only 14% of the descriptive snippets we extracted and analysed with SnipIt). Juru was then tweaked to support snippets and to receive a separate boost for texts of that type.

We ran Juru with the new retrieval unit as an integral part. We first measured the binary effect of the snippets – with and without them. We then gradually increased the influence (boost) of the snippets on the overall ranking score. For each run we computed the official evaluation measures using TREC's eval program.

5.2.2. Results

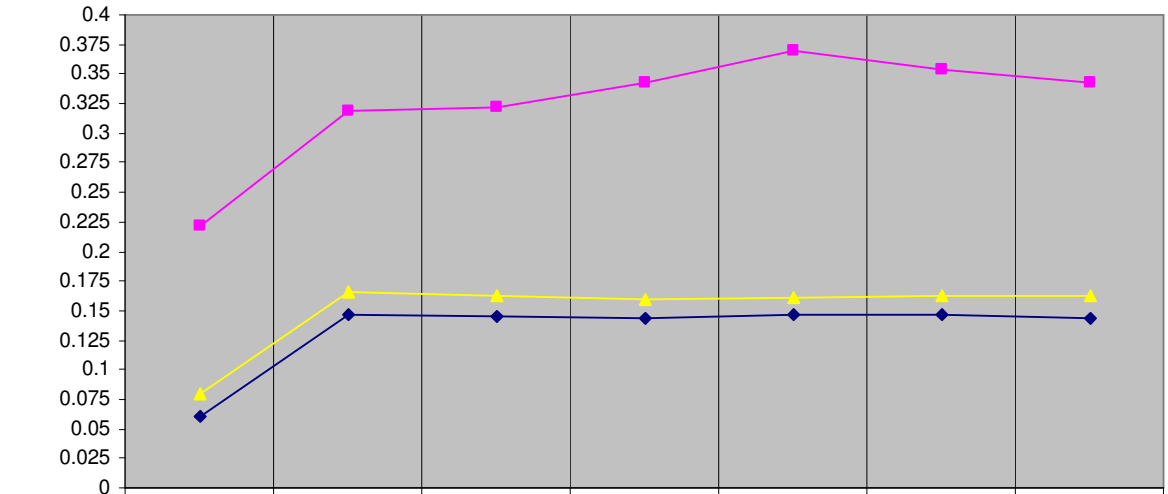
Results for the set of experiments are summarised in Figure 9 and Figure 10. Using snippets as the only searchable element returned correct answers in the top 20 results for 70% of the queries. Using only text and anchors as searchable units achieved a corresponding success rate of 87%, and using all the units combined achieved a success rate of 88%. When we increased the boost of the snippets we were able to raise the success rate to 90% (i.e. we were able to bring at least one correct answer in our top 20 results).



	text/anchors = 0 snippets = 1	text/anchors = 1 snippets = 0	text/anchors = 1 snippets = 1	text/anchors = 1 snippets = 5	text/anchors = 1 snippets = 10	text/anchors = 1 snippets = 15	text/anchors = 1 snippets = 20
P@1	0.157	0.218	0.227	0.267	0.3	0.267	0.256
P@5	0.091	0.175	0.175	0.164	0.167	0.167	0.158
P@10	0.083	0.148	0.147	0.144	0.141	0.143	0.136
P@15	0.075	0.134	0.133	0.134	0.134	0.135	0.13
P@20	0.077	0.124	0.122	0.124	0.122	0.123	0.121

Figure 9 A summary of the 7 different runs of Juru with descriptive snippets as a stand-alone retrieval unit. The boosts were assigned in 5 units increments, as they appear in the ranking function itself.

Figure 9 shows the increase of precision@1 (i.e. the correctness of the first ranked result) as we increase the boost of the descriptive snippet text, achieving its maximum improvement around *boost = 10*. This fits well with the intuition that snippets have some correlation with PageRank. Specifically for a topic distillation task, such as the one practiced in TREC, it seems that the contribution of the descriptive snippet as a retrieval unit has an affect on the first ranked results. Prec@5, Prec@10, Prec@15 and Prec@20 remained almost without change, and were not deemed affected by the introduction of the new retrieval unit. Figure 10 strengthens this notion of improvement at the highest ranked results by showing a peak around *snippet boost = 10* for Mean Reciprocal Rank (MRR), showing virtually no change in Mean Average Precision (MAP) and R-Precision.



	text/anchors = 0 snippets = 1	text/anchors = 1 snippets = 0	text/anchors = 1 snippets = 1	text/anchors = 1 snippets = 5	text/anchors = 1 snippets = 10	text/anchors = 1 snippets = 15	text/anchors = 1 snippets = 20
MAP	0.061	0.146	0.145	0.143	0.147	0.146	0.143
MRR	0.222	0.319	0.322	0.342	0.369	0.354	0.342
R-Prec	0.08	0.166	0.163	0.16	0.161	0.162	0.163

Figure 10 A summary for the 7 different runs of their Mean Average Precision (MAP), the Mean Reciprocal Rank (MRR), and the R-Precision (precision after all relevant results for a given query were retrieved)

5.2.3. Discussion

We believe our results show a greater correlation between popularity measures and the existence of descriptive snippets. We attested the contribution of snippets for the top-most ranked results and decreased the amount of noise associated with taking text around the anchor as a separate indexing unit. We believe treating descriptive snippets as separate from the anchor text is crucial for their use as retrieval units of text. The fact that someone made an attentive effort to describe another Web page within their text is a sign that the added text is mandatory for explaining that target page within the context of the pointer page. In topic distillation tasks this is precisely the kind of meta-information we are looking for. By storing and indexing descriptive snippets separately we allow our search system to "reward" and boost such descriptions. Adding this extra layer of information on top of the anchor layer may enhance systems using anchor text to search for tasks such as homepage finding [10][17].

It may be that as stated in [20], the .GOV collection is biased towards a well connected centre. It is also possible that the 100 queries used in both topic distillation tracks of TREC 2002 & 2003 gravitated as well toward this centre. It will be interesting to test this hypothesis on the larger collection to be collected for future TRECs.

6. Conclusion

This paper presented a system that uses layout regularities to automatically detect and extract paragraph-anchor patterns. The tool developed, SnipIt, takes advantage of the observed paragraph convention found in Web hypertext. It automatically extracts descriptions and other scribbles that people write about other texts, and assembles them. The vision behind the tool described is similar to the one presented by Landow ([17]) in his chapter "Reconfiguring the Author". He envisions another 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)

SnipIt, and the concept behind it, aim to show that there is a relatively simple and inexpensive way to process human language on the Web while preserving some characteristics that are lost when applying less linguistically motivated tools. We hope to see SnipIt implemented in Web search engines on a regular basis in order to improve the crawling, indexing, and ranking outputs, and also to improve the ability to summarise the content of indexed pages. The only aspect we did not tackle so far is a system for building automatic directories of Web pages that would benefit from having SnipIt as an automatic description identifier. There is no need for costly maintenance. SnipIt provides a constant tracking mechanism for updating stale and outdated descriptions. Such

an application will reuse descriptive snippets for automatically clustering and categorising Web sites based on people's descriptions ([7][24]). This is the last frontier for SnipIt's exploration.

7. Acknowledgements

The Royal "We" used in this paper is certainly not a stylistic choice. It is used to stress the fact that I have been fortunate to share my enthusiasm for text around anchors with many great minds and amiable people. I would like to thank some of those good people without whom SnipIt was merely a nonsense word. To Cecile Paris and Michael Johnson for their wise and friendly supervision; to Sarah Boyd for the name SnipIt and for moral support; to Steve Green for his faith in Perl and me; to Monika Henzinger for taking charge; to Martin Colton-Farach for deciphering my Perl and thoughts and for being in charge of Google's WebQuotes application; to Rani Nelken who implemented SnipIt in WebFountain; to David Carmel & Ronny Lempel for making space and assigning power to snippets within Juru; and to Uri Parker for enduring this obsession of mine for so long (and counting…).

8. References

[1] 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. http://www.webir.org/resources/phd/Amitay_MSc_1997.pdf

[2] Amitay E. (1999). Anchors in context. in Words on the Web - Computer Mediated Communication, Lynn Pemberton & Simon Shurville eds., Intellect Books, UK.

[3] Amitay E., Paris C. (2000). Automatically Summarising Web Sites - Is There A Way Around It? ACM 9th International Conference on Information and Knowledge Management (CIKM), pp. 173-179.

[4] Amitay E. (2001). Trends, Fashions, Patterns, Norms, Conventions... and Hypertext Too. Journal of the American Society for Information Science (JASIS). (Special Issue on Information Science at the Millennium), 52(1):36-43.

[5] Amitay E. (2001). What Lays in the Layout: Using anchor-paragraph arrangements to extract descriptions of Web documents. PhD Thesis, Macquarie University, February, 2001. http://www.webir.org/resources/phd/Amitay_2001/index.html

[6] Amitay E., Carmel D., Darlow A., Herscovici M., Lempel R., Soffer A., Kraft R., Zien J. (2003). Juru at TREC 2003 - Topic Distillation using Query-Sensitive Tuning and Cohesiveness Filtering. NIST Special Publication SP 500-255, The Twelfth Text Retrieval Conference (TREC 2003).

[7] Attardi G., Di Marco S., Salvi D. (1998). Categorisation by Context. Journal of Universal Computer Science (J.UCS), Springer Verlag, 4:9:719-736.

[8] Bailey, P., Craswell, N., Hawking, D. (2003). Engineering a multi-purpose test collection for web retrieval experiments. Information Processing and Management, 39(6):853-871.

[9] Brin S. and Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the Seventh World Wide Web Conference (WWW7), Brisbane, also in a special issue of the journal Computer Networks and ISDN Systems, Volume 30, issues 1-7, pp. 107-117.

[10] Craswell N., Hawking D., Robertson S. (2001). Effective site finding using link anchor information. ACM SIGIR 2001, pp. 250-257.

[11] Chakrabarti S., Dom B., Gibson D., Kleinberg J., Raghavan P., Rajagopalan S. (1998) Automatic resource compilation by analysing hyperlink structure and associated text. Proceedings of the 7th International World Wide Web Conference, Brisbane, also in a special issue of the journal Computer Networks and ISDN Systems, Volume 30, issues 1-7.

[12] Clark H.H., Wilkes-Gibbs D. (1986). Referring as a collaborative process. Cognition, 22, 1-39. Also reprinted in P. R. Cohen, J. L. Morgan, & M. E. Pollack (Eds.), Intentions in Communication. Cambridge: MIT Press, 1990.

[13] Davison B.D. (2000). Topical Locality in the Web. ACM SIGIR '00, pp. 272-279.

[14] Delort J.Y., Bouchon-Meunier B., Rifqi M. (2003). Enhanced web document summarization using hyperlinks. ACM Hypertext 2003, pp. 208-215.

[15] Eiron N., McCurley K.S. (2003). Analysis of anchor text for web search. ACM SIGIR 2003, pp. 459-460.

[16] Haas S.W. & Grams E.S. (2000). Readers, authors, and page structure: A discussion of four questions arising from a content analysis of Web pages. Journal of the American Society for Information Science (JASIS), 51, 2, 181-192.

[17] Hawking D., Upstill T., Craswell N. (2004). Toward better weighting of anchors. ACM SIGIR 2004, pp. 512-513.

[18] Landow G.P. (1991). Hypertext: The Convergence of Contemporary Critical Theory and Technology. Johns Hopkins University Press, 1991.

[19] Lorch, R.F., Jr., & Lorch, E.P. (1996). Effects of headings on text recall and summarization. Contemporary Educational Psychology, 21:3:261-278.

[20] Soboroff I. (2002). Do TREC web collections look like the web? ACM SIGIR Forum, 36(2):23-31.

[21] WayBack Machine of the Internet Archive, http://www.waybackmachine.org

[22] WebFountain framework for data mining, http://www.almaden.ibm.com/webfountain

[23] Google WebQuotes, http://labs.google.com/cgi-bin/webquotes

[24] Zamir O.E. (1999). Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results. PhD Thesis, University of Washington, 1999.

[25] Zhang Y., Milios E., Zincir-Heywood N. (2004). A Comparison of Keyword- and Keyterm-based Methods for Automatic Web Site Summarization. In Technical Report WS-04-01, AAAI'04 Workshop on Adaptive Text Extraction and Mining, pages 15-20, San Jose, CA, USA, July 26, 2004.