Overview of text summarization in the context of information retrieval and interpretation: Applications for web pages summarization

Álvaro Mendes Barbosa
Pompeu Fabra University - Audiovisual Institute
Estació de França Paseig de Cincurvalacio, 8 Desp. 391
08003 Barcelona, España – http://www.iua.upf.es
Tel.: +34 93 542 2588, email: alvaro.barbosa@tecn.upf.es

The World Wide Web as content source

The World Wide Web is today the largest repository of information available, and increasingly is becoming the main source of content for many different forms of expression in modern society.

The content of the web includes different kinds of audiovisual information (text, pictures, video and audio), however the majority of the available information is still in textual form i.e. coded in natural language.

In this sense it is not surprising that recent research and proposals for new projects focused on web information retrieval, mostly concern textual information, even if the final goal is to address other kinds of data usually trough the use of metadata.

On the other hand with the common paradigms for retrieving and interpretation of web content by a regular user, there is a great degree of inefficiency regarding the amount of information that the user as retrieved until he can actually interpret it in an meaningful way.

This is not a new problem in the information society. Taking for example a common case as the information display on a newspaper, we found headlines and carefully created titles that are inserted strategically so that the reader can get a general overview of the contents before he decides if he wants to read more in-depth information.

It is clear that in modern information retrieving and interpretation understanding and comprehension are roles that more and more are being displaced into to the information provider then into the receiver, with the objective of reducing the latency time between information retrieving and interpretation.

In this context the summarization of textual content in an automatic way is a major research area, and the development of systems that are oriented to the specificities of the web information are object of great interest in the computer science community.

The web characteristics

Retrieving information from the web has a common background with classical information retrieval mostly based on databases of structured text documents, however a typical web document has specific characteristics that need to be taken into account.

With the developments in network technology at a very low cost the remote access to various sources of information became worldwide spread, however with this accessibility also came issues related multicultural background implying users with different ways to perceive information and specifically the issue of multilingual content become of great importance.

With web technology also came the freedom of publishing information for the common user even tough he might not create an organized structure and the published content may lack relevance and clearness.

Al this makes the typical content within a web document non-coherent and with an structure and language that is not clearly defined, often resembling a chaotic jumble of text phrases, links, graphics and formatting commands [2].

On the other hand, the concept of Hypertext and the way the web it self is structured has also a great influence on the way users retrieve information.

No one really knows what is the structure of the World Wide Web, since it grows faster than it's own ability to detect its changes, in a way that the interconnectivity between web pages is dynamic and a big percentage of these connections became obsolete and are never updated.

The way in which the information is distributed in the web over different computers and different platforms added up to the fact that the great majority of the pages that exist are generated automatically reinforces that the approach to perform any summarization task based on web documents will necessarily have to be automatic.

We can conclude that the web is a highly interactive medium much due to its hyperlinked structure, but still the most effective and widely used ways to find information is based on traditional information retrieval. New techniques adapted to the medium need to be developed.

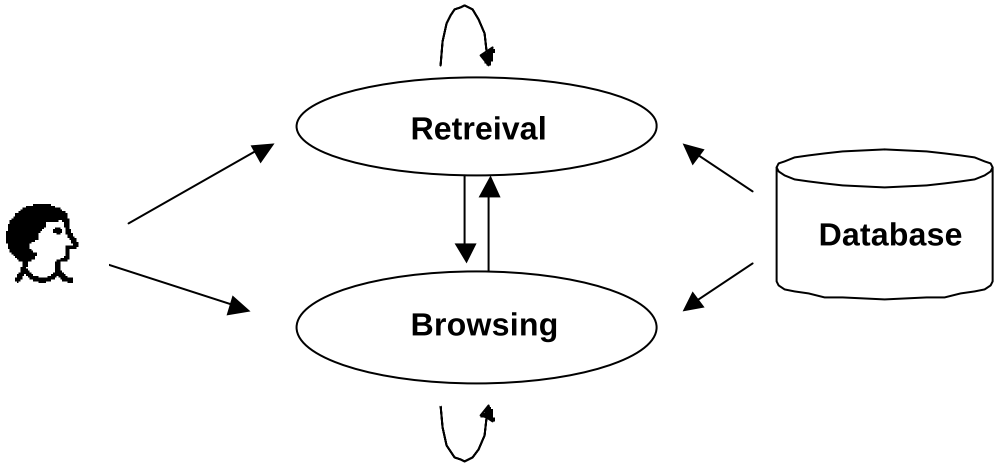
In order do develop these new techniques issues like quality of indexing and better interaction systems for the user should be address and in this context automatic summarization of web pages has an important role.

The user task [1]

It is a common approach when analyzing the users role to make a distinction between the tasks he or she might be engaged.

When using a data retrieval system the user needs to translate the criteria he defined in order to find the desired information into a language that is supported by the information retrieval system. In order to do so the most common ways are to express the user's criteria with a query an there fore he is performing a **Searching Task**, or if the systems supports it, the user can conduct a **Browsing Task** trough collections of documents with a general idea (that can change in the course of the work) and locate the desired information.

Classical information retrieval systems allow information or data retrieval and Hypertext systems are usually oriented to enhance browsing capabilities, however the combination between retrieval and browsing is not yet a well established and is not the dominant paradigm even over web interfaces.



The previous figure illustrates the interaction of the users with a database of documents (such as the web) through the tasks of browsing and retrieval.

Both browsing and retrieval are classified as pulling actions, in the sense that they are a literal response driven by the user request for information in an interactive fashion.

Enhancements in the functionality and effectiveness in the process of information retrieval can be achieved with techniques to push information towards the user.

These techniques usually aim to provide additional useful information to the user in the context of is requests, and therefore achieving more efficiency.

In this sense, providing summary's extracted from the content that the user is browsing in an automatic fashion is one of the most obvious and effective ways to perform this task.

However there is a distinction to be made that often is misleading between information extraction and information summarization from a text in a given document or a set of documents.

The differences between information extraction and information summarization lie mainly in the techniques used to identify the relevant information and in the ways that information is delivered to the user.

Information Extraction is the process of identifying relevant information where the criteria for relevance are predefined by the user, while in text summarization the user does not necessarily start with a predefined set of criteria or interest, but is the system that automatically should determine which information will be relevant for the user.

Text Summarization

Summarization is one of the most important capabilities required in writing.

Performing a summarization task in an automatic fashion is by it self a big challenge and has been object of major research over the 5 decades.

A short history of references on this field was presented by the professor of computational linguistics Udo Hahn's from the University of Freiburg on his presentation at Automatic Summarization Workshop in Seattle (2000).

Early Extraction 1955 - 1973

- Luhn, 1955, lexical occuence stats
- Edmundson, 1969, positional indicators, lexical cues, cuewords, cue phrases
- Mathais, 1973, cohesion streamlining

Linguistic Approaches 1961 - 1979

Psychological Approaches 1975 - 1985

Artificial Intelligence Approaches 1980's

- 1982. DeJong. FRUMP, using scripts
- 1985. SUSY, logic and production rules
- 1988. Reiner and Hahn.TOPIC. Frames and semantic networks
- 1989. Rau et al. Hybrid Representations

"Renaissance" - 1990's

- Recurrence of statistical techniques
- Hybrid approaches

In order to generate automatically a summary from a text there are typically two paths one can follow. One can built the summary upon the extraction of contiguous, coherent spans of text from the document, in which case we are performing **Extractive Summarization**, or one can use **Non-Extractive Summarization** building up the summary in a meaningful way using words or collocations (linguistic expressions that have meaning as a hole, like adjective-noun phrases, such as “information retrieval”) that some how are more representative in the text.

Much work has been done in extractive summarization in particular during the Tipster **[5]** program of research and development in the areas of information retrieval, extraction, and summarization, funded by DARPA and various other North American and European Government agencies from 1991 to 1998.

The third phase of the program was partially focused on the issue of extractive text summarization also addressing multilingual issues and the resulting developments were deployed within the scientific community to provide analysis with improved operational tools. Due to lack of funding, the program formally ended in the Fall of 1998.

However extractive summarization has several drawbacks **[6]** including the inability to generate effective summaries shorter than a sentence, whish is problematic when it is required a short “headline” style summaries with only a few words, since sentences with summary content are usually longer than average, and information in the document is often scattered across multiple sentences.

Besides extractive summarization cannot combine concepts in different text spans of the source document without using the whole spans.

Furthermore due to the unstructured, disjointed and non-coherent characteristics of **web documents**, little foothold is provided for extractive summarization techniques since the required spans for extraction within the text need to be contiguous and coherent.

In fact most important prior work in extractive summarization has explored aspects that rely in properties of the source text that web pages often lack and therefore an efficient summarization cannot be performed with this technique.

Summarization of Web Pages

Recent research activity have addressed the ambitious goal of non-extractive summarization oriented for web pages, being that the most significant results where presented in the SIGIR’00 – the Special Interest Group in Information Retrieval conference in Athens, Greece (2000), by Adam L. Berger and Vibhu O. Mittal in the paper **OCELOT**: A system for summarizing web pages **[2]**.

The OCELOT system's approach is to synthesize a summary rather than extract one, relying on a set of statistical models to guide it's choice of words, and how to arrange these words in a summary.

The statistical models are built using standard machine learning algorithms, which takes as its input a large collection of human summarized web pages obtained from the Open Directory Project [7].

The Ocelot project bears close relation to recent work in automatic translation of natural language, in the sense that in this case statistical models are built using machine learning algorithms with human translations between two different languages as an input, to generate a statistical automatic translation, however, the degree of difficulty in the case of text summarization is much higher, since a satisfactory translation of a sentence must capture it's entire meaning, while a satisfactory summary is actually expected to leave out most of the source document's content.

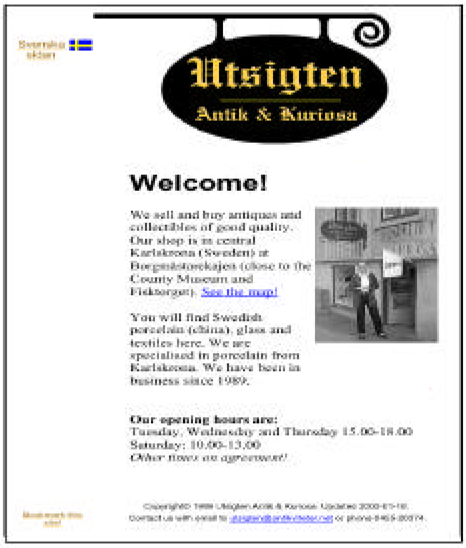
In order to be able to apply machine learning to this problem using web pages and human-summaries as an input, the content of the web page had to be treated so that it could be handled by the system, in a way that resembles the typical indexing procedure when creating a text database for an information retrieval system.

The following steps where applied to the original database of web documents and human-generated summaries (Open Directory Gists) pairs:

- Text Normalization: punctuation was removed; all text was converted to lowercase; all numbers were replaced by the symbol NUM; each occurrence of the 100 most common overall words where removed (stopword-filtering).
- All Links, Images and Meta information where remove from the web pages
- All web pages containing adult-oriented content were removed.
- All HTML markup information was removed
- Pairs whose web pages contained frames were removed
- Pairs whose web pages were just "page not found errors" were removed
- Pairs whose web pages or gists were too short – less than 400 or 60 characters respectively
- All duplicate web pages were removed
- The remaining set of pairs was partitioned into a train set containing 99% of the pairs and a test set containing 1% of the pairs.

After processing the average length of the summaries was 13.6 words and the average length of a document was 211.1 words.

The following figure shows a "before and after" example of this filtering process on a single web page along with the Open Directory's gist.



Svenska sidan

Utsigten

Antik & Kuriosa

Welcome!

We sell and buy antiques and collectibles of good quality. Our shop is in central Karlskrona (Sweden) at Borgmästarekajen (close to the County Museum and Fisktorget). See the map!

You will find Swedish porcelain (china), glass and textiles here. We are specialized in porcelain from Karlskrona. We have been in business since 1989.

Our opening hours are:
Tuesday, Wednesday and Thursday 15.00-18.00
Saturday: 10.00-13.00
Other times on agreement!

Bookmark this site!

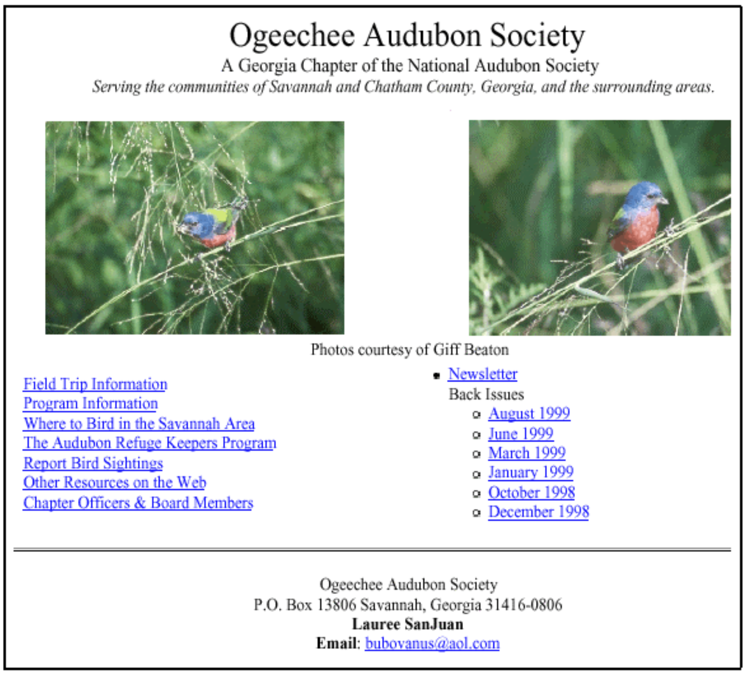
Copyright© 1999 Utsigten Antik & Kuriosa. Updated 2000-01-10.
Contact us with email to utsigten@antikviteter.net or phone 0455-20374

Filtered: svenska sidan utsigten antik kuriosa welcome we sell and buy antiques and collectibles of good quality our shop is in central karlskrona sweden at borgmstarekajen close to the county museum and fisktorget see the map you will find swedish porcelain china glass and textiles here we are specialized in porcelain from karlskrona we have been in business since NUM welcome to our shop our opening hours are tuesday wednesday and thursday NUM NUM NUM NUM saturday NUM NUM NUM NUM other times on agreement bookmark this site copyright NUM utsigten antik kuriosa updated NUM NUM NUM contact us with email to utsigtenantikviteter net or phone NUM NUM

Open Directory gist: sell and buy antiques and collectibles of good quality our shop is in central karlskrona sweden

These data was used to train the statistical models, and subsequent tests where performed on the 1% of web pages and OD gists pairs that were not observed in the learning process.

The following figures show two examples of the OCELOT's behavior over the evaluation set.



Ogeechee Audubon Society

A Georgia Chapter of the National Audubon Society

Serving the communities of Savannah and Chatham County, Georgia, and the surrounding areas.

Photos courtesy of Giff Beaton

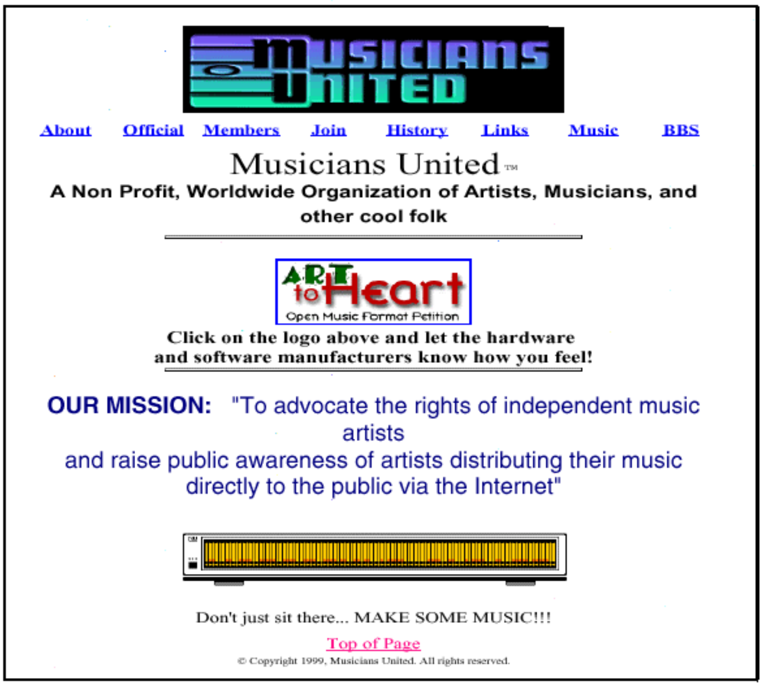
Field Trip Information
Program Information
Where to Bird in the Savannah Area
The Audubon Refuge Keepers Program
Report Bird Sightings
Other Resources on the Web
Chapter Officers & Board Members

- Newsletter
  Back Issues
  - August 1999
  - June 1999
  - March 1999
  - January 1999
  - October 1998
  - December 1998

Ogeechee Audubon Society
P.O. Box 13806 Savannah, Georgia 31416-0806
Lauree SanJuan
Email: bubovanus@aol.com

Open Directory gist: a chapter of the national audubon society serving the communities of savannah chatham county and the surrounding areas

OCELOT *gist:* audubon society atlanta area savannah georgia chatham and local birding savannah keepers chapter of the audubon georgia and leasing



About Official Members Join History Links Music BBS

Musicians United™

A Non Profit, Worldwide Organization of Artists, Musicians, and other cool folk

Art to Heart
Open Music Format Petition

Click on the logo above and let the hardware and software manufacturers know how you feel!

OUR MISSION: "To advocate the rights of independent music artists and raise public awareness of artists distributing their music directly to the public via the Internet"

Don't just sit there... MAKE SOME MUSIC!!!

Top of Page

© Copyright 1999, Musicians United. All rights reserved.

Open Directory gist: to advocate the rights of independent music artists and raise public awareness of artists distributing their music directly to the public via the internet

OCELOT *gist:* the music business and industry artists raise awareness rock and jazz

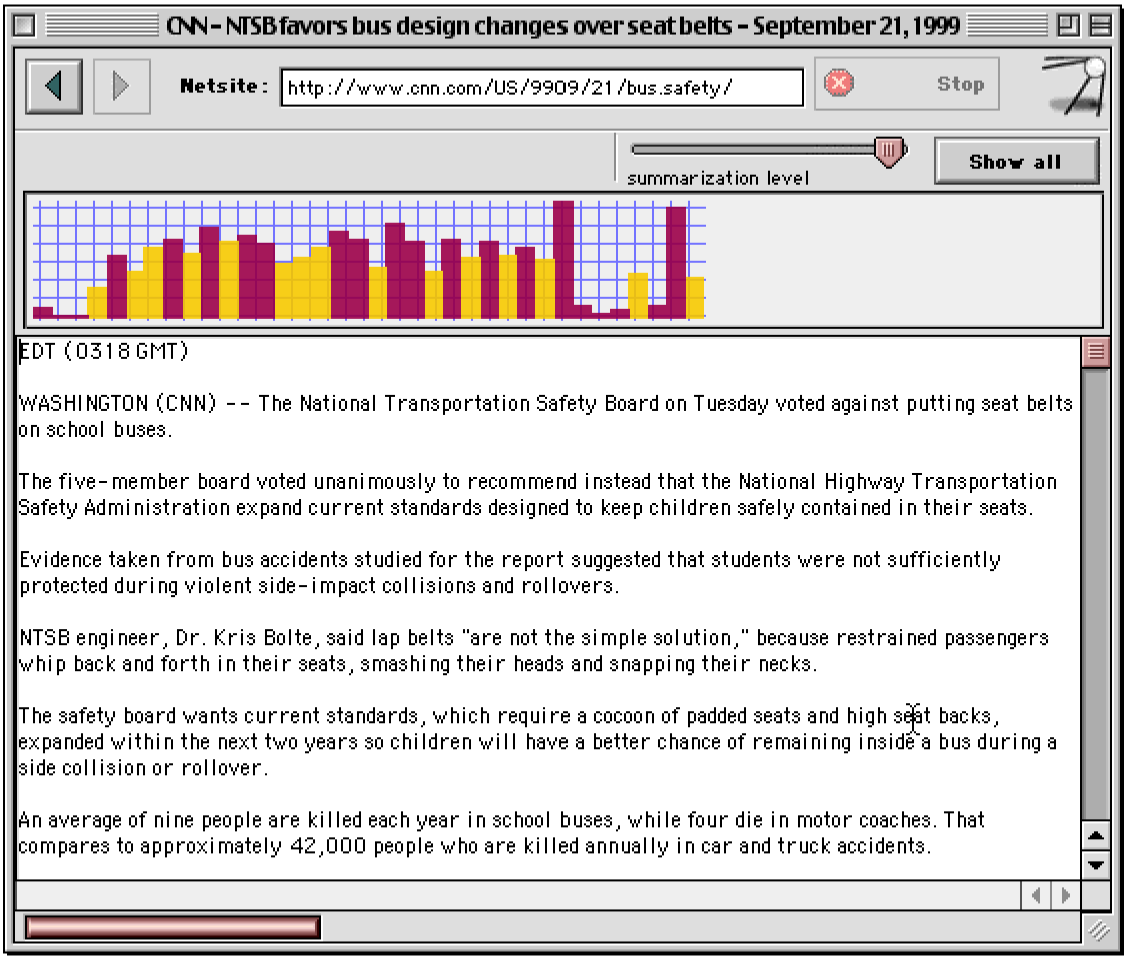
As a concept the OCELOT system is also suitable for translingual summarization (automatically summarize a web page written in a language different from the generated gist) as long as one can use pars of web documents and human-generated summaries written in different languages as the input to the machine-learning algorithm.

Besides the OCELOT system research project there are also preliminary efforts to create commercial products that propose automatic summarization of web pages, specifically the **Data Hammer** software developed by Glucose Development Corporation [8] is a relatively successful product.

The Data Hammer system consists of a stand-alone application for the Macintosh platform on whish the user can drag and drop links from a web browser and it will perform a summary extraction from the text contained in the respective web page.

However Data Hammer is based on extractive summarization and it is only suitable for well structured documents like for instance news articles that can be found in the web, and in this sense although it is an application that can be used with web pages it is not an application oriented to the summarization of the normal textual content existing in the web.

In the product documentation [9] is claimed the characteristic of the software being scalable, but it only concerns the fact that the user can interactively define parameters of the summarization like for instance the minimum length required for a sentence to be selected has suitable to integrate the summary.



CNN - NTSB favors bus design changes over seat belts - September 21, 1999
Netsite:
http://www.cnn.com/US/9909/21/bus.safety/
Stop
summarization level
Show all

EDT (0318 GMT)

WASHINGTON (CNN) -- The National Transportation Safety Board on Tuesday voted against putting seat belts on school buses.

The five-member board voted unanimously to recommend instead that the National Highway Transportation Safety Administration expand current standards designed to keep children safely contained in their seats.

Evidence taken from bus accidents studied for the report suggested that students were not sufficiently protected during violent side-impact collisions and rollovers.

NTSB engineer, Dr. Kris Bolte, said lap belts "are not the simple solution," because restrained passengers whip back and forth in their seats, smashing their heads and snapping their necks.

The safety board wants current standards, which require a cocoon of padded seats and high seat backs, expanded within the next two years so children will have a better chance of remaining inside a bus during a side collision or rollover.

An average of nine people are killed each year in school buses, while four die in motor coaches. That compares to approximately 42,000 people who are killed annually in car and truck accidents.

The previous figure is a snapshot of the data-Hammer interface, where one can observe in the graphic the yellow bars as a representation of the important sentences, therefore available to be selected as part of the summary and the purple bars as the sentences that are less important and that wont be part of the summary.

Using hyperlinks in summarization of Web pages

In the notion of Hypertext are introduced the concept of Node and link as its basic components [12].

A node is document or a fragment of a document, like one static html page, and a link is a logical connection between nodes.

Hyperlinks are obviously the key syntactical element for hypermedia navigation over the web structure, however they also have a semantic value that can be exploited.

Even dough the semantics of web links are not expressed explicitly it is likely that it could be taken into account when trying to create a summary out of a web document (the OCELOT system simply removes all links before performing any analysis), although this might require the parsing of inter-connected pages.

The concept of using the semantic context of a link is not new in web information retrieval, in particular when indexing databases for search engines, where many web documents are indexed by the link text and are not fiscally retrieved to the local server, being that documents indexed this way can even have non html formats, like post script or pdf.

The semantic value of a text span that links a web object is also often used to index pointers to non-textual information like in the Image search engines from AltaVista or Google [13].

It is intuitive to realize that if a web page author decides to create a link out of a text span then it probably means it's content has relevance in the overall context of the document, however it is not so clear weather it is relevant for a summary that describes the essence of the document.

In fact many times links are added to documents to point out information that is not of much relevance to the current theme, but to offer the user a possibility to pursue a different theme in another document that could be totally unrelated to the main context.

In some specific cases it is more clear if the text from a link should have incremented relevance when performing a summary, like for instance in the case of a link within the same document (an anchor).

In this case it is clear that there is a very high probability that the link sentence somehow represents one section of the document, and there fore is more relevant.

On the other hand, for instance links that correspond to on-site navigation, clearly wont have much relevance to a summary.

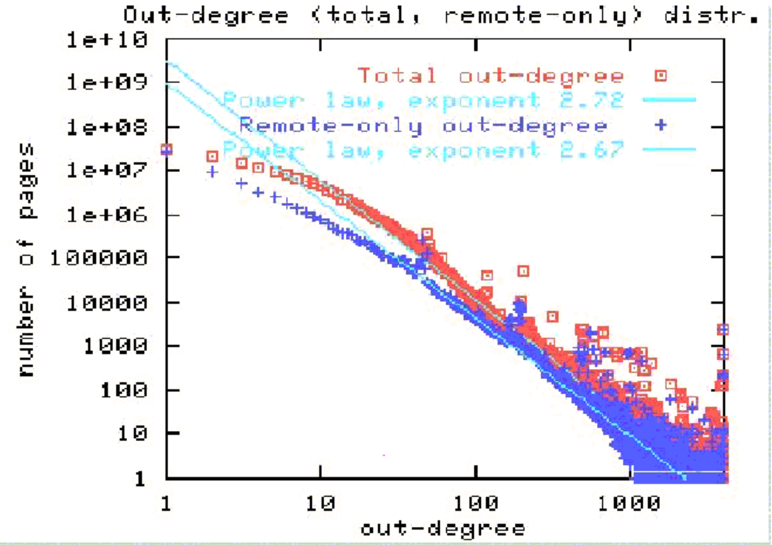
Links can also be explicit when build at authoring time or implicit when build at access time, and since it is a known fact that the majority of the existing web pages are generated automatically, also this should be taken into account.

When trying to build a summary out of a web page, the relevance of an explicit link and an implicit link is not the same, since when the process is driven by human cognition capabilities such as in the case of explicit links it would most likely be more relevant in a process of summarization that tries to mimic the human achievement of reducing textual information to its most essential points.

Considering a node has a flat web document and following an approach based on statistical models built upon machine learning algorithms that could analyze the way a human generated summary would use information from both explicit and implicit links, some knowledge could be provided in weather the textual content of an hyperlink should be ranked above the average when building a summary automatically.

However the greatest potential that comes from taking hyperlinks in account when creating a summary, is the possibility of referring to the content of the linked nodes.

From the analysis of degree of distribution of out links from single web page one realizes that the great majority of existing pages in the web has a limited amount of out links (up to 10) **[14]**.



Pages with higher amounts of out links would most likely correspond to a hub page (response from a search engine qwery or web page created manually with a list of links to a particular subject).

In a hub page links are pointed mostly to pages with redundant information about the specific subject of the page, so it is reasonable to think that in this pages one would easily found more meaningful information to create a summary, however this would be difficult to compute due the extensive amount of links that this pages can have, and on the other hand this are not the most common type of existing pages.

On pages with smaller amounts of out links (that correspond to the majority of the existing web pages) the degree of uncertainty on the relevance to a summary of “second generation” information within linked pages, might increase due to the lack of content to analyze, but on the other hand if one consider that in this case the links would be more often created manually and therefore based on human cognition, the relevance could be Higher.

In any case in order to the use of information from linked pages in a summary, it is hard to evaluate relevance due to the high degree of uncertainty in the relation with the original text, however this could be minimized by:

- Parsing the original web page in such a way that one could determine explicit and implicit links.
- Determine if the page is a Hub.
- Perusing the analysis on these pages only on a second iteration so that the relevance of its content would be proportional to the rank of the link's text within the main page.

It is also clear that non-extractive summarization techniques should be used on the linked pages, since usually there is also no structural relation between the main page and the linked page, especially if this pages are located in remote servers.

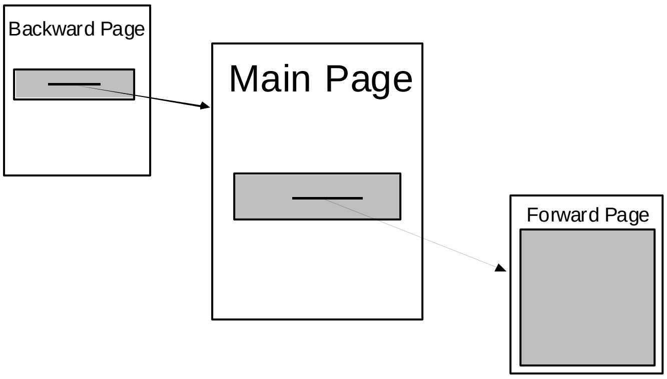
Another possibility to explore the potential of hyperlinks is to approach a web page from its backward links.

Early work published in 1996 at the fifth World Wide Web conference [15], unveiled the potential of backward navigation as an improvement current navigation paradigm in the web based on following forward links.

From an extensive catalog of web pages it can be easily determined which web pages have links to a specific web document, and in fact search engines like Google and AltaVista provide this possibility.

The motivation to look for web pages that point to specific web document is quite clear, since it is very likely that a page that refer to a given document ought to be very valuable for the detection of related information.

Further more it is very likely that the backward link text is very relevant to the summary of the current page, since a link text is usually done in such a way that by reading it the user can have an idea of the content of linked the page.



It is reasonable to assume that, in fact the content of a section from a web page that points to the web page that is being summarized is more likely to be useful in the summarization process than the content of a page pointed by an out link from the page to be summarized, since has a concept whenever a link is inserted in on text section it usually points to a page that has more information on a specific theme being approached in the section.

This concept is more applicable for explicit than implicit links, however it might be general enough to be used with a non extractive summarization approach that also accepts text spans an text links from pages that point to the main page being summarized.

Although the use of hyperlinks and its semantic value in the process of summarization of a web page seams to have great potential, the biggest drawback is that in order to implement a system that uses some of the presented ideas, it would be necessary to have accesses to a big enough collection of web pages with respective manually made summary in order to obtain statistical models that could be used on automatic summarization.

Conclusions

It is clear that to unveil the full potential of web information retrieval, interpretation and summarization it is necessary to use different techniques than the ones traditionally used in information retrieval systems due to the specificities of web documents and to the structure of the web itself.

In particular for text summarization very few work that takes in account the specificities of the web has been published by the scientific community, therefore it is an open research field of great importance to modern society.

Based on my understanding of a High-level overview to the current state of the art in this field I would like to point out some of the direction that I believe will be pursuit in the near future in order to produce better summarization systems that are appropriate to the web paradigm:

- Mechanisms to deal with ambiguity and better use of the semantic context are necessary especially in non-extractive summarization. For instance in the OCELOT system it is not possible to distinguish “Dog bites Man” from “Man bites Dog”

- In some specific applications it can be performed a query oriented summarization or at least it should be taken in account previous knowledge about the target user when performing the summarization

- Multilingual summarization is one of the main issues to be addressed

- More information regarding the structural clues embedded in a web document from the HTML or XML code should be exploited. A simple example is for instance that it is more likely for a word to be relevant to a summary if it is included within the Tag <TITLE>…</TITLE> then if it is within the tag <SMALL>…</SMALL>

- It should also be taken in account the structure of the web documents and the way hyperlinks work. In particular the semantic value from linked text and within linked web pages can also be exploited, especially in the backward links domain (links from pages that point to the web page being summarized).

References:

[1] Modern Information Retrieval, Ricardo Baeza-Yates y Berthier Ribeiro-Neto, Addison-Wesley, Wokingham, England, March 1999.

[2] OCELOT:A system for summarizing web pages, Adam L. Berger, Vibhu O. Mitta. In Proceedings of SIGIR'00. Athens, Greece (2000).

[3] Recuperación de la información: Algoritmos, Estructuras de datos y búsqueda en la web, Gonzalo Navarro, Ricardo Baeza-Yates, Universidade de Chile 1999.

[4] Measuring the Web, Tim Bray, Proceedings from the fifth World Wide Web conference. Paris, France (1996)

[5] TIPSTER Text Program A multi-agency, multi-contractor program:
http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/

[6] Ultra-Summarization: A statistical Approach to Generating Highly Condensed Non-Extractive Summaries, Michael J. Witbrock, Vibhu O. Mittal, from the proceedings of Sigir'99.

[7] The Open Directory Project – a collection of approximately 900.000 web pages each with a short annotated human-authored summary
http://dmoz.org/

[8] Glucose Development Corporation - http://www.glu.com/

[9] MTT: A scalable text summarization engine suitable for embedding within consumer products, 1997 - Glucose Development Corporation

[10] Project ISIR – Italian Spanish Information Retrieval final report, Dipartimento e Informatica – Università di Padova; Departamento Matemàtiques I Informàtica – Universitat de les Illes Baleares

[11] Summarization Resources website, Stephen Wan - *Macquarie University – Sydney Australia* - http://www.ics.mq.edu.au/~swan/summarization/

[12] Class notes for the course Artificial Intelligence and Information Retrieval on the PhD program on Computer Science and Digital Communication – Universitat Pompeu Fabra, Barcelona Spain, 2001 - Massimo Meluci.

[13] Google image Search: *http://image.google.com/* ; Altavista Image Search: *http://www.altavista.com/sites/search/simage/*

[14] Networks and Sub-Networks in the World-Wide Web, Prabhakar Raghavan, Verity Corporation, 2000.

[15] Web Core - Forward and Backward browsing in a Multi-grain Web representation. Carlos Baquero, Jorge Portugal: Distributed Systems Dep. Informatics - Minho University, Portugal - Poster Session at the Fifth International World Wide Web Conference, 1996 - Paris.