Web Sites Automatic Summarization

© Mikhail Kondratyev

Saint-Petersburg State University
mikhail@oasis.apmath.spbu.ru

Abstract

Summarizing web pages and web sites is a challenging problem. While research in the area of automatic summarization has almost 50 years long history there are many open questions and web sites summarization is one of them.
In this paper we represent preliminary results of our research of automatic web sites summarization methods. The goal of this work is to identify the limits of applicability of context-based approaches for web sites.

1 Introduction

Typical web site is rather big object containing lots of information. In many situations it is handy to be able to describe its content in few lines. Examples of such descriptions are web site entries in Web directory (such as dmoz.org) or site descriptions in results produced by a search engine.

At the moment the only way to produce a good summary of web site content is to use human to do this manually. Most of Web catalogs follow this approach. However, it is obviously impossible to summarize each site in each search engine response manually. To be able to cope with high volumes of information available today in the Web it is crucial to have automatic techniques which can produce meaningful summaries.

Currently there are lots of techniques developed for text documents summarization with research started in this area several decades ago. Fast Internet growth stimulated research in web content summarization but there are still few approaches known for web sites summarization task. The traditional techniques can't be directly applied to web documents because of the following reasons:

- **Web documents are not isotropic**. Most of Web pages contain different media, such as text, pictures, sound, flash movies, etc. This media may be very important for human users but is very difficult to summarize automatically. In fact, a web page may contain no text content at all.
- **Web pages may contain no direct information**. Indeed, the purpose of many web pages is not to give information on a subject in the way of textual (video, sound, etc.) description but to reference other web documents. Good sample of such pages are big portals' main pages, such as www.msn.com or www.yahoo.com main pages.
- **Part of information may need to be ignored**. Web pages often contain lots of irrelevant information that suites navigation, design or advertisement purposes. Extracting relevant information is a separate task that may heavily depend on the user's intentions.
- **One logical document may be presented on several pages**. Sometimes a document may be divided into several parts with each part presented on a separate page. Such pages usually link to each other and retrieving the whole document requires page's HTML structure analysis.

To solve these problems a set of context based approaches was suggested ([1], [3]). These approaches use web document's context to extract information for automatically generated summaries. Under the document's context we mean content of all the web pages linking to this document. The basic assumption here is that links often contain a textual description of the target page somewhere around the anchor.

Although modern web documents summarization techniques extensively use hyperlinks information and web content's specific features such as documents tree structure, there are several additional challenges that need to be addressed when summarizing web sites:

- **Web sites are complicated objects**. Sites usually contain lots of web pages, with both static and dynamic content.
- **Pages are not equally important**. Some of them contain no information at all to be added to the summary. Good example is a company's online web catalog – it may contain lots of textual descriptions but they contain no information about the site itself.
- **A single site may cover several topics**. Such topics may be completely independent (like weather and business on the yahoo site) or just cover different areas of a company's business (for

Proceedings of the Spring Young Researcher's Colloquium on Database and Information Systems SYRCoDIS, St.-Petersburg, Russia, 2005

example software, hardware, services, online documentation, etc.). All these topics may need to be summarized.
- **Web sites may differ significantly**. There exists no generic pattern to be applied to all sites.

In this paper we introduce our context-based approach to web sites summarization. The approach was developed as part of web sites summarization research with the purpose of investigation of applicability and effectiveness of several context-based algorithms and heuristic rules. To work around difficulties of web site summarization, the proposed algorithm uses web structure information and analyses the pages referring to the target web site to find summaries somewhere in the context. Supposing that there should be a description of the site under question somewhere in the Web, we try to find such readymade descriptions and select the one that suites our purposes best.

2 State of the art

The automatic summarization research is about five decades old with lots of different summarization algorithms developed for wide range of summarization tasks. In this section we give a brief overview of the main trends in research area.

Summarization tasks differ one from another which results in various types of the summaries. A summary, for example, may be user-focused, i.e. tailored to the requirements of a particular user or user-group or may be generic. In [6] the following summary types are marked out:

- Detail: indicative/informative
- Granularity: specific events/overview
- Technique: extraction/abstraction
- Content: generalized/query-based
- Approach: domain/genre specific/independent

Various algorithms and strategies are used for summary generation. The traditional automatic summarization techniques are mainly based on the content of the document and therefore may be called content-based summarization techniques. There are three major groups in this area of research. First of them is paragraph based summarization group. Algorithms belonging to this group try to divide text documents into paragraphs and extract the sentence (-s) that describes the paragraph in the best way (the first sentence in the paragraph for example). Such approaches are usually used for big textual documents covering lots of different topics.

The second group of approaches tries to identify most informative sentences from the document based on one or more rules and then assemble them together. The resulting summary likely will not contain coherent text but should be as informative as possible. Such algorithms may analyze terms frequencies, sentences positions in the text (headers, paragraph starting sentences) or apply other techniques.

The approaches from the third group use natural language processing cues in the text. Such cues may include order of words, acronyms identification, synonyms and special phrases. The heuristics used in such approaches tend to be domain-specific and may require special tuning when summarizing documents from different topic domains.

Another important problem is evaluation techniques which can be used to compare different approaches. Several researches have shown (see [4] for example) that human judgments may poorly overlap. Unfortunately, currently there are no standard techniques for summaries quality evaluation as the properties to be measured and summary types vary significantly. The example of a very important but difficult to measure property is readability. Informativeness is also difficult to measure automatically in case of abstracted summaries. A popular metric for extractive summaries was proposed by Edmundson in [2]. Using this metric generated summaries are evaluated by calculating precision and recall of sentences (or terms) presented in both automatically created and human made extractions.

3 Our approach

3.1 Web sites summarization

Web site summarization is a challenging task that is very difficult to perform automatically. To workaround these challenges we assume that there should be a human made site description somewhere in the Web. Indeed, there are lots of site descriptions in the web directories, on user home pages, etc. Such descriptions usually represent an expert's opinion in a coherent and well readable way. Consider a big web portal, such as www.yahoo.com. Such web sites usually contain information related to the great range of categories, including business, weather, news, web search, email services, and others. Automatically generating a summary for such site will likely result in the set of non-coherent sentences describing various parts of the web portal. Human-made summary, on the contrary, presents information in readable and coherent way:

The first large scale directory of the Internet, now a major portal offering search engine results, customizable content, chatrooms, free e-mail, clubs, and pager (taken from www.dmoz.org)

We use web structure and web pages analysis to extract readymade descriptions, annotations or summaries created by other people that describe the target site. We assume that there should be a suitable description somewhere in the web and our task is to find and extract such descriptions automatically for the given site.

Our approach follows the same idea as the 'In common sense system' described in [3] but uses different rules for summaries search and selection. We also aim to provide Russian language specific algorithms while all the known researches in this area use English-specific heuristics. Some other works based on the idea described in [3] suggested to mix sentences from different web pages (see [1] for example) but this one adds the problem of sentences coherence.

In addition to the coherence problem it brings up the problem of including too specific sentences in the summary. Consider the following reference found:

www.myco.com *– The MyCo, announces new line of scalable AMD x64 servers targeting enterprise customers market.*

Such reference may align well with the company's business but it doesn't contain information about the company itself and describes just a single activity.

The task of producing summaries following our approach requires the following problems to be solved:

- How to find pages referencing the target site?
- What text should be extracted from the link context?
- What are the rules for the best summary candidate selection?

Summaries evaluation should be done to check how effective the suggested approach is.

3.2 Finding pages from the context

In our research we did not implement any searching mechanisms but decided to use well known search engines such as google (www.google.com) or yandex (www.yandex.ru). Currently we use google's web services based API to access its functionality. Google was selected for search purposes because it is known as one of the best search engines and provides easy-to-use API for its services although other search engines may be used as well.

In [3] queries of type *link:www.targetsite.com* are suggested but experiments have shown that such queries return too many irrelevant pages and too small percent of them contains good summary candidates. This was especially clear when querying for smaller sites when we often could not extract a single candidate. Queries of type *' "*www.targetsite.com *" '* behaved a bit better but still did not satisfy our needs.

The initial search result has great impact on the overall system quality so a set of experiments was done to enhance the query. The experiments have shown that excluding links from the site to itself helps to filter out lots of pages that just link to the site but don't describe it (see topicality section). This is especially useful when summarizing big sites as the search engine often returns their pages. Such pages usually link to the main page as part of navigation system only and don't have any descriptive information in the links' context.

As most of site descriptions may be found in different web directories and catalogs (such as www.dmoz.org or www.yahoo.com) adding keywords like 'directory' helps to improve search result quality significantly.

The final queries we used during approach evaluation match the following pattern:
<keyword> "<target site>" –site:<target site domain> where *<keyword>* is 'directory' or, for example, 'directory OR catalog', *<target site>* is the site to be summarized, *-site:<target site domain>* allows to exclude pages from the target site. It is clear that such queries may be constructed automatically based on the site address. The samples of queries are:

*directory "*www.sun.com*" –site:sun.com*
or
каталог OR директория www.rbc.ru *–site:rbc.ru*

3.3 Extracting text

When extracting text snippets from the pages referencing a web site we have to solve topicality and partiality problems [1]. Topicality is the problem of web pages referencing the target site but containing no information about the site itself. Consider the following example : *"<a href="www.cnn.com">CNN</a> has reported the results of today's elections"*. Text fragments may also contain information about company business activities but not the site itself which also makes them irrelevant to Web site summarization task, for example*"<a href="mycompany.com">My Company</a> has reported lower then expected profit in the second fiscal quarter*". Partiality is the problem of context documents describing only part of the target site but not the whole one. Our experiments have shown that in general more than 90 percent of the pages retrieved by a search engine have to be rejected due to the partiality or topicality issues. In our approach we use two steps algorithm to solve partiality and topicality problems.

In the first step initial filtering is performed and in the second step the final summary is selected. Initial filtering allows rejecting most of irrelevant pages linking to the target site. According to the experiments less than 15 percent of summary candidates left after the initial filtering contains noisy data and can't be used as a summary.

The first step is based on the common rules that web page authors usually use when describing a site. The [3], for example, suggests extracting text fragments of the following form only:
<paragraph start><anchor><text without links><paragraph end>

Taking this rule as a basis we tried to soften it a bit and tried also to use the following patterns:
<paragraph start><anchor><text><paragraph end>
and
<sentence start><anchor><text><paragraph end>,
where text may contain other links. Our experiments have shown that using the last pattern generates too many irrelevant text fragments which makes it less effective than the former one. The experiments also helped us to identify the set of rules for paragraphs extraction. We define paragraph here as, like in [3], any text fragment visually separated from the other content. To identify a paragraph's border a special set of 'border' tags was created.

According to our results most of text fragments fitting into the text extraction pattern contain enough descriptive information to be a good candidate to become a summary. The important observation is that although applying the pattern is critical for generating summaries it can do nothing if the initial search query returned mostly noisy information

3.4 Selecting the best summary candidate

This step is the most complicated one. The step consists of applying different rules to the extracted text snippets to assign them a score. Each rule may assign negative or positive scores depending on the snippet content, context or based on the whole snippets set analysis. The problem of balancing the scores being assigned still needs to be solved. The current values for scores were adjusted manually based on our experiments. In general, doing this manually may result in wrong total scores assigned to candidates and may significantly increase number of selection mistakes.

Our experiments helped us to identify the following problems with the summary candidates left after initial filtering:

- Summary candidates may contain no coherent text.
- Summary candidates may contain too little data.
- Summary candidates may represent someone's personal opinion like *'www.myworstsite.com is the worst site I have ever seen'*.
- Summary candidate may refer to the target site but describe something else, for example *'www.thatcompany.com was accused of breaking about 100 US Patents recently. The problem of intellectual property is now widely discussed in Europe and the United States. Experts say that creating serious software tools without breaking a patent is almost impossible now. Big companies tend to patent every invention however unimportant it is while smaller companies can't afford this'.*

This paper represents preliminary results only and we hope to significantly improve candidate selection algorithms and extend the rule's set. We started from implementation of several rules that measure summary quality based on different text snippet properties. All rules that require language-specific knowledge were implemented for Russian language only. The most important are stop word quantity rule, term frequency rule, personal opinion rule and size rule. We consider these rules in the sections below.

Size rule checks a candidate size. The size is calculated as the number of terms of the summary candidate. When applied, the rule compares a candidate's size with default minimum threshold and assigns negative score in case the size is below minimum. Our experiments show that bigger summaries usually contain more descriptive information and rarely contain noise data. Assuming this we grant bigger summaries with higher scores.

The stop word rule's aim is filtering out summaries containing too many stop words and special signs like '|', '=', etc. Big number of such terms usually means that the extracted text snippet is a part of navigation system or just overlaps with another description. We defined 33 percent as the maximum ratio in our experiments. All candidates that have more than 33 percent of stop words (from the total count) were assigned negative score by this rule.

The personal opinion rule was created to solve the problem of personal opinion descriptions. Such descriptions may refer to the target site but contain someone's opinion instead of expert knowledge of the subject. To find out such summaries the stemmed terms are compared with special vocabulary containing words usually used to express someone's opinion. We assume that even if one such word exists it means the summary should be assigned a negative score.

The term frequency rule's aim is analyzing the frequency of terms in summary candidates. The idea behind this rule is that all descriptions of the same object should contain common words, while those text snippets that have very little in common with the majority will likely describe something else and were selected by mistake. Consider there were selected some candidates with the same text. In case we don't filter out the repeated candidates the rule will assign higher scores to them. To avoid this the recurring summary candidates should be removed from the set. The problem here is that such candidates often are not absolutely the same and may differ just in one or two terms. To solve the problem we decided to use the *Jaccard similarity* measure for summaries comparison, defined as

$$Sim = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are strings represented as term sets. Calculating this measure and comparing it with the predefined threshold allows avoiding the similar candidates in the candidate set. Term frequency for the term *w* was calculated as

$$Tf = \frac{\sum_{A \in S} C_{w,A}}{\sum_{A \in S} |A|}$$

where $C_{w,A}$ is the number of occurrences of term *w* in summary *A* and S is the set containing all summary candidates. A candidate was assigned higher scores in case it contains most frequently used terms.

Our research has shown that most of the 'good' summary candidates were extracted from various online web directories and link lists. Web sites descriptions may be found on home pages as well as in huge web directories like *dmoz* and may differ significantly in structure. Although such summary sources may look completely different there is usually a set of records on their pages with all records having similar structure. Existence of such records in the page is a good clue and may be a reason for assigning higher scores for summaries extracted from this page. The rule should try to identify the similar records on the page and check if the summary candidate is one of such records. This rule is still under development.

4 Evaluation

4.1 Evaluation results

The aim of evaluation was investigating how our approach and implemented heuristics work, what parts need to be improved and what should be paid additional attention in the further research. Due to current implementation limitations and time constraints the evaluation was done based on the set of 30 Russian web sites randomly selected from the top-level categories of www.list.ru catalog. For each of the sites about 80 pages from the site context were analyzed and text snippets extracted. Not more than 20 summary candidates were selected for a site. After applying the rules the best candidates were chosen based on their score. The summaries selected were compared to the hand made summaries of the www.list.ru catalog.

Assessors were asked to evaluate the produced summaries. Each summary was judged based on the good, bad, excellent and unknown marks scale. Good mark was assigned when summary was acceptable as the site's summary, bad mark in the opposite case. Excellent mark was assigned to the best ‘good’ candidate. The unknown mark was assigned when a summary was not readable.

The experiments have shown that at least one candidate was found for about 80 percent of the sites. Those sites that could not be found following our approach seem to be little known in the Russian part of the Web and have too small context. Increasing number of sites that we can produce summaries for is an unsettled question. One possible way to solve it is modifying initial search query. It is also possible that more than 80 referring pages should be checked for summary candidates to increase the number of sites with summaries found.

The evaluation results have shown that heuristics used for candidate selection require more tuning as too many summaries were assigned the same score. This made almost impossible evaluating how good the prototype was in selecting the best summary candidate. Anyhow, only about 15 percent of the candidates retrieved following the defined rules were marked with 'bad' mark (which means 85 percent of the candidates may be used as a summary). Unfortunately there were technical problems with current implementation working with Russian encoding which lead to relatively big number of 'unknown' marks (about 20 percent).

All numbers above were calculated counting unique summaries only. The following table contains an example of evaluation results.

Site	*Candidate*	*Mark*
www.telur.ru	Telur.ru - интернет-магазин бытовой техники Ассортимент: аудио-, видео- и фототехника, бытовая техника, электроинструменты. Цены. Условия доставки. www.telur.ru Цитируемость:750. Регион: Москва	good
	Каталог товаров с иллюстрациями и характеристиками: бытовая техника, ТВ и видеотехника, аудио и Hi-Fi, фототовары, презентационное оборудование, климатическая техника, автоэлектроника, средства связи, игровые приставки. Цены. Информация об оптовых поставках и транспортно-экспедиторских услугах.	excellent
	TELUR.ru - Интернет магазин бытовойтехники и электроники. Аудио. Видео. ТВ. Телефония. Климатическая и Бытовая техника.Всегда на складе более 3000 наименований. Лучшие цены. Доставка по всей России.Партнерская программа.	good
	TELUR.Ru - интернет-магазин аудио, видео, ТВ и бытовой техники Аудио, видео, ТВ, бытовая техника и электроника оптом и в розницу по лучшим ценам с доставкой по всей России. www.telur.ru	good
	http://www.telur.ru/ - Совместное российско-испанское предприятие "ТЕЛУР" http://www.telta.perm.ru/ - ОАО Пермский телефонный завод "ТЕЛТА" kamatel.perm.ru - Камател - цифровые системы передачи информации http://www.skif.permonline.ru/ -	bad

Site	*Candidate*	*Mark*
	Пермский завод "Машиностроитель" http://www.permenergo.ru/ - ОАО "Пермэнерго" http://www.prometey.perm.ru/ - завод им. Кирова torgmash.perm.ru Пермский завод торгового машиностроения	
	ОПЧПУФЙ ПФ TELUR оПЧПУФЙ ЛПНРБОЙЙ TELUR, ФПЧБТЩ Й ГЕОЩ - ОПЧЩЕ РПУФХРМЕОЙС, ФЕНБФЙЮЕУЛЙЕ ПВЪПТЩ, БОБМЙЪ ТЩОЛБ ВЩФПЧПК ФЕИОЙЛЙ LG Electronics... http://www.telur.ru	unknown

Our experiments have shown that summaries with the same content or with minimal content differences were often extracted from various sources. This shows that big part of the summaries is copied from one, the most authoritative, source. Finding this source may be useful for better summaries scoring: summary candidates retrieved from the authoritative site may be assigned higher scores. One of the possible ways to determine such trustworthy summary sources (assuming the source is a web directory) is intersecting sets of equal summaries source sites:

$$Sres = St_1 \cap St_2 \cap St_3 \cap \ldots \cap St_n$$

where St_n is the set of source sites containing equal summaries for the target site t_n. The resulting set will contain source sites that host repeated summaries for the web sites t_1, t_2, t_3 … t_n. Such source sites are potential candidates to be the most authoritative ones.

4.2 Problems found

The experiments held helped us to identify several problems to be solved in the future.

The time needed for summarizing a web site can't be predicted and may significantly differ from site to site and even from time to time for the same site. This happens due to various web pages download speed and timeouts when page is not accessible. To overcome this we plan to implement multithreaded web documents download.

There is relatively big group of sites that have smaller context and can't be summarized following our approach. This group includes newly created sites, small private web sites and others. For example, the SYRCoDIS site could not be summarized using our approach because Google could find only 2 links to it. We could also observe the situation when only one unique summary candidate was found for a site.

Page encoding is determined now using HTTP headers only but many documents do not contain this information. Encoding support should be extended by reading HTML *meta* tags info or by analyzing term/letter usage statistics and comparing it with known statistics for the Russian language.

5 Conclusion

In this paper we present the preliminary results of our web sites automatic summarization research. We had described context-based approach that uses information around links to given site to produce site summary. The evaluation results show that in general this approach seems promising but there is number of problems identified to be addressed in the further research.
At the moment our approach relies on the set of heuristic rules that probably are not the best possible. We plan to investigate if we can learn better rules with machine learning techniques.

References

[1] Delort, J.-Y., Bouchon-Meunier B., Rifqi M. Enhanced Web Documents Summarization Using Hyperlinks. *ACM*, 2003.

[2] Edmundson, H.P., New Methods in Automatic Extracting. In *Journal of the ACM*, 1969. 16(2): p. 264-285.

[3] Einat Amitay, Cecile Paris. Automatically Summarizing Web Sites – Is There A Way Around It? *CIKM 2000,* 2000.

[4] Goldstein, J., V. Mittal and J. Carbonell. Creating and Evaluating Multi-Document Sentence Extract Summaries. In CIKM'00: *Ninth International Conference on Information Knowledge Management*. 2000.

[5] Inderjeet Mani. Summarization Evaluation: An Overview. In *proceedings of the NTCIR Workshop*, 2001.

[6] Inderjeet Mani. Automatic Summarization*,* John Benjamin’s Publishing Company, 2001.

[7] Madhavi K. Relevance of Cluster size in MMR based Summarizer: A Report. *Document Understanding Conference,* 2002.

[8] Simone Teufel, Marc Moens. Sentence Extraction as a Classification Rask. In *ACL/EACL*,1997

[9] William W. Cohen, Pradeep Ravikumar, Stephen E. Fienberg. A Comparison of String Distance Metrics for Name-Matching tasks. *American Association of Artificial Intelligence*, 2003.

[10] DMOZ Web site. www.dmoz.org

[11] Google Web site. www.google.com

[12] List.ru Web site. www.list.ru

[13] Yahoo Web site. www.yahoo.com