Conceptual Text Generation Based on Key Phrases

M. Charnine\textsuperscript{1}, N. Somin\textsuperscript{1}, and V. Nikolaev\textsuperscript{1}
\textsuperscript{1}Institute for Informatics Problems of the Russian Academy of Sciences, Moscow, Russia

Abstract - The method and system for automatic generating meaningful articles called Conceptual Texts from key-phrases found on the Internet is presented. Conceptual Texts are intended to describe basic concepts of subject domain and their relationships. Key features of Keywen system for generating Conceptual Texts is discussed. Shown how to select meaningful key-phrases and estimate their informativeness. Ranking approach to automatic estimation of informativeness of phrases and documents is presented. The definition, an example and the purpose of creation of Conceptual Texts are presented. General classification of Internet texts including Conceptual Texts is presented in the form of two-dimensional diagram with coordinates of originality-informativeness. Methods of description of subject domain by a list of keywords and documents are discussed. Structure of further research is suggested.

Keywords: Conceptual Texts, key-phrases, originality, informativeness, Internet, Big Data.

1 Introduction

Rapid growth of information in the Internet gives us the opportunity to find relevant documents about many interesting topics. Each topic can be defined by a set of keywords, or key-phrases, or by a set of relevant documents. There are millions of different keywords and phrases, and the number of their subsets (topics) is more than the number of people on the Earth. Most of these topics don’t have ready up-to-date encyclopedic articles with description of basic relevant concepts and their relationships. We call such articles as Conceptual Texts. Conceptual Text is a natural language text containing complete and accurate review of basic concepts of theme and their relationships. This work is presented a software tool called Keywen for generating Conceptual Texts or Conceptual articles. If you define some topic by a set of keywords then Keywen tool will find relevant key-phrases from Internet. These key-phrases can be used as building blocks for generating Conceptual Texts. The tool also can automate the process of compiling Conceptual Texts from these key-phrases. The Conceptual Texts are useful source of ideas that can help in writing scientific and encyclopedia articles and reviews.

From a user point of view the process of generating Conceptual Texts consists of the following steps. The user:

- creates the overall design and conceptual schema of article on a given topic,
- formulates the list of related keywords,
- investigates texts and phrases found by the tool in terms of the task,
- selects most relevant and informative phrases (key-phrases),
- organizes them according to the conceptual schema.

Below is a general classification of the Internet texts from the point of view of informativeness and originality. We also show the location of Conceptual Texts in this classification.

2 Classification of Internet texts

Let us consider the classification of Internet texts in terms of their information content (informativeness) and originality. We also distinguish computer-generated texts and texts written by people.

In [1] the authors propose to consider the structure of Internet texts presented by a two-dimensional diagram with coordinates of originality and informativeness.
In this approach different types of texts occupy different zones in this diagram, and so can be successfully identified based on these two parameters.

According to the degree of the originality the texts can be divided into original texts, plagiarism and severe plagiarism. Original author's texts do not contain incorrect citations. These texts have a maximum originality and informativeness (meaningfulness) and therefore they are placed in the upper right corner of the diagram.

In plagiarism there are incorrect citations of external ideas without proper references. In severe plagiarism there are incorrect citations of large chunks of someone else's texts without proper references.

Based on informativeness the Internet texts can be divided into informative and meaningful author's texts, computer-generated meaningful texts and computer-generated meaningless texts. Author’s Conceptual Texts have a maximum informativeness and are placed at the top of the diagram. Computer-generated Conceptual Texts are meaningful texts and are placed in the middle of the diagram.

Lower left corner of the diagram should be occupied by uninformative and meaningless plagiarism. Such texts in practice don’t exist because usually only informative and meaningful texts are plagiarized.

This article is devoted to methods of generating Conceptual Texts, as well as related issues of determining informative texts.

3 Conceptual Texts: the purpose of creation

Since Conceptual Texts are created for reading by people, the main characteristics of Conceptual Texts are their meaningfulness and informativeness. Under the informativeness we understand the ability to carry information that is perceived by readers as meaningful message in the form of natural language text. The Conceptual Texts are substantially different from so-called "generated noise" that includes all sorts of texts not intended for reading by people, but intended for processing by computer programs.

Note that, in contrast to the "generated noise", the generated Conceptual Texts are usually not intended for misrepresentation (although such use is not impossible). Their primary goal is to serve quite legitimated information purposes. The rapid growth of number of documents on the Internet makes it possible to find not only relevant, but also the most quality information. The fact is that, along with informative and accurate documents there are a huge amount of poor quality information, misinformation and disinformation with unknown authorships. And now the users are not satisfied by finding documents related to the question of interest, because some of these documents can be of poor quality. Now the problem is formulated differently: select from a given subject domain the best possible information that has not only relevance, but also has high degree of accuracy, i.e. select the information that is verified and can be accepted by the scientific society. Such quality information from the subject domain can be collected in the form of Conceptual Texts. Conceptual Texts can be created manually by people or generated by automatic programs.

Below is an example of a manually created Conceptual Text with the title “How to understand something”.

DEFINITIONS

Understanding is a process that mirrors the world.

To understand something is to know that person or concept deeply.

To understand something means to be able to describe it in terms that are intuitive.

SEE

If you would like to understand something, you should not hear but see it.

The best way to understand something is to see it with your own eyes.

PARTS

To understand something as a whole, you need to understand the parts that go into.

QUESTIONS

To be able to understand something we must for a start be able to answer six questions about it: WHAT?, WHY?, HOW?, WHEN?, WHERE?, WHO?

STUPID THING

The only stupid thing you can do is pretend to understand something when you don't.

QUICKLY

The best way to understand something quickly is to read Keywen Encyclopedia.

This Conceptual Text/article was created by M. Charnine from short phrases found on the Internet. The article is interesting to readers, contains definitions and basic concepts related to “understanding” ("see", parts, questions). Also this article contains examples/descriptions of relationships
between basic concepts in the form of natural language phrases.

It is possible to automatically generate Conceptual Texts by using automatic analysis of texts from the Internet. Automatically generated Conceptual Texts can be used as starting point for creating different kinds of automatic encyclopedias and reference books. This approach is used for the development of the program Keywen [2, 11] intended to build this kind of encyclopedias.

Internet encyclopedia Keywen.com contains over 250,000 articles and exists on the Internet for over 10 years. This is the first encyclopedia in which users can contribute by electronic voting for the best category, keywords or phrases. Over the past three years over 120,000 users have voted for the best definitions and phrases and became co-authors of encyclopedia.

Below, we describe an approach for creating systems like Keywen that are able to generate Conceptual Texts.

4 Ranking approach for generating Conceptual Texts

At a first sight it may be concluded that the task of generating Conceptual Texts of sufficient quality is a very difficult problem of artificial intelligence. Indeed, the automatic Conceptual Texts generator must solve two challenges. First, the generator should create some meaningful content (ideas) that is interesting and useful for readers. And secondly, the generator should present this meaningful content in the form of natural language text that looks like a "normal" and "readable" document. Both of these tasks are close to the simulation of human thinking process of writing of high quality texts. The solution of both tasks is now in the embryonic stage, so it is difficult to expect that artificial intelligence methods soon will be able to solve these problems of generating Conceptual Texts.

However, we can go another way. We can compile Conceptual Texts from small fragments (or phrases) that are already available in the Internet. We can choose the most relevant and informative fragments. But here the following question arises: how to evaluate the quality of information content of these text fragments? To address this issue, we use a following hypothesis: “the most informative phrases are often quoted and rephrased”. In other words, the Conceptual Texts will be of better quality if they are composed from the phrases presented ideas that are popular in the Internet. Such an approach we call ranking approach, meaning that the oft-cited phrases (and respectively - documents) have high citation index and a high rank in search engine results.

It should be noted that the ranking approach is far from universal. There are some documents and phrases that are very valuable and meaningful, but for whatever reason they did not get a high citation index. For example, an article which sets completely new questions and contains new terminology after publication will definitely have low citation index, although its information content can be of a very high quality. However, it will take some time during which the scientific community will appreciate new ideas of the article, "vote" for them and increase citation index, or “vote” against them, leaving the article without links. This means that the ranking approach estimates informativeness only in static, but not takes into account the dynamics of information processes.

Another limitation for the ranking approach is the risk of false ratings. It is not a secret that currently many websites use special "promotional" technologies and incorrect "black" schemes (such as "doorways") which significantly distorts the true rating. That is why the ranking approach should be accompanied by the tools for identification of meaningless documents to avoid their influence on the final rating. So, we still need to use some of artificial intelligence technologies. But this kind of problems can be already solved today.

5 Description of subject domain

It is not a trivial task to create description that characterizes domain boundaries. In Keywen the subject domain is defined by a set of keywords/terms. These keywords can be individual words or phrases. Later this set of keywords is used to compile the search queries. This method is sufficiently universal and is convenient for the users.

In [3] the authors present another way to describe the subject domain by using a set of links to documents that adequately describe the subject domain. Thus, the domain is defined by a set of texts/documents. An objective comparison of the two methods of defining domain (by keywords and by documents) was not done yet. It can be assumed that the method of defining domain by a list of documents gives less informational noise. However, sometimes it is more difficult for the user to create a list of documents (in this case the system itself can create the list of related documents).

6 Method for generating Conceptual Texts

In Keywen [2, 11] a search query is generated based on a set of keywords. The algorithm for generating search query is iterating through all combinations of keywords (single keywords, pairs of keywords, triplets, etc.) so that the query length does not exceed a certain threshold N. The threshold N
is determined empirically so that search results contain several hundred of relevant documents in average.

Then the search queries are processed using search engines (Google, Yandex and other), and as a result a set M1 of text documents is retrieved. The set M1 can be used as a source of URL-links, and as a result the extended set of documents M2 is formed. A set M2 also contains URL-links, and this way the process of expanding the set of documents M2 can be continued by several iterations of additions of linked documents. It is necessary to check the newly found documents for the presence of the original keywords. After subsequent iterations the set of required documents M2 can be significantly increased.

Next, the set of documents M2 is divided into text fragments (or phrases) similar in length to the conventional sentences. As a result we have a database with records in the form

<text fragment> - <URL, indicating this fragment>

The size of this database can be up to several terabytes. To store and search in the database it is not enough traditional software tools working with databases, but it is necessary to use special hardware and software technologies named «Big Data».

By using the obtained database the statistical calculations are made: how many different independent URLs refer to the same text fragment and its semantic equivalents (paraphrases and translations). We calculate semantic equivalents using automatic translator and linguistic processor [5].

Note that there is a problem of independence of web pages/documents. The fact is that for advertising purposes there are often created doubles of pages, which content almost completely duplicates the content of the main web page. We should delete such doubles from statistical calculations. We can identify possible doubles of web pages and web sites by using our accumulated database, because it actually stores all the important content from all web sites.

Thus, each text fragment gets a rating equal to the number of occurrences of its semantic equivalents in the independent web pages. The top rated fragments (key-phrases) are used for generating Conceptual Texts of a given subject domain.

We believe that in most cases this approach can provide the perfectly acceptable result, since the Internet publications of scientific articles in open access became more and more popular. This is partly due to the fact shown by Algirdas Ausra [4], that such works are cited much more actively than those available by subscription or at the library.

7 Rating of informativeness

The proposed ranking method for generating Conceptual Texts can be used with some modifications for evaluation of the informativeness of Internet documents. Indeed, if we calculate the total weight for each document as the sum of ratings of all its text fragments, then this total weight can be considered as an estimation of informativeness. Note that this estimation is subject to all the comments that we have made for Conceptual Texts generation.

The estimation of the originality of Conceptual Texts obtained by the ranking method is not a simple problem. On the one hand, Conceptual Texts are original, because of the unique combination of its fragments (key-phrases). But if we consider the document up to fragments, it turns out that the document is composed of text fragments (like from bricks) taken from the Internet. Thus, the estimation of the originality of the document depends on the algorithm, so the systems for identifying plagiarism don’t have a clear and simple solution.

8 Conclusions

This paper presents a method and system for automatic generating meaningful articles called Conceptual Texts from key-phrases found on the Internet. Conceptual Texts are intended to describe basic concepts of subject domain and their rrelationships. Conceptual Texts are created for reading by people, and the main characteristics of Conceptual Texts are their meaningfulness and informativeness. Conceptual Texts are original, because of the unique combination of its fragments (key-phrases). The proposed ranking method for generating Conceptual Texts can be used with some modifications for evaluation of the informativeness of Internet documents.

Further research is expected in the following directions:

• enhanced methods for detection and monitoring of the most important concepts of subject domain to describe them in the Conceptual Text [6, 8, 9];

• methods for generating the order of the key-phrases and the structure of sections of Conceptual Texts [11];

• enhanced methods for discovering informative key-phrases on the Internet in accordance with their paraphrases and translations to other languages [10];

• special linguistic processors for discovering semantic similarity of phrases expressing the same idea [5, 10].
9 Acknowledgements

This work is supported by the Russian Foundation for Basic Research, grant #13-07-00272 “The methods for automatic creation of associative portraits of subject domains on the basis of big natural language texts for knowledge extraction systems”.

10 References


