elpub2002 Proceedings — J. A. Carvalho et. al. (Eds.) — VWF Berlin, 2002

Intellectual Information Technologies and Scientific Electronic Publishing

Changing World and Changing Models

Mikhail G. Kreines

Moscow Medical Academy
Moscow Center for New Information Technologies
in Medical Education
2-6, B. Pirogovskaya Str., Moscow, 119992, Russia
Tel.: 07-(095) 2458624, 07-(095) 9673343
http://www.mmascience.ru; kreines@mmascience.ru

Abstract. We discuss new information technologies of semantic analysis, search and retrieval of textual information and their influence on the world of electronic publishing. We shell discuss the means and tools for semantic analysis of texts and texts collections for information search and retrieval, for detection of novelty in texts corpora and their usage in a lot of applications. New structures and new models of electronic publishing for efficient solving of existing problems in the fields of information quality, Copy right and effective ways and means of information retrieval are under consideration.

In the proposed models the large and authoritative collections of the texts are real bases for the efficient solving of existing problems. So, the owners of the large information collections – universities, research centers and publishers have the key position.

The intellectual data analyze uses this collections to make right solutions of the both readers' and writers' problems.

1 Introduction

It is a common knowledge that our world forces people to change technologies and models of business and behavior. And changed technologies and models dramatically influence the world itself. In this paper we discuss the possible evolution and problems of scientific electronic publishing due to alterations of the world of the scientific information and information technologies. New models and roles in the world of scientific publishing are under consideration.

Gutenberg had really made democratic revolution in access to the texts as carries of the scientific and social knowledge and cultural heritage. After Gutenberg it became just a financial problem to get any printed edition. In general publishing has grown into successful business. New types of the texts' carries (as journals and newspapers) had appeared.

Nowadays publishing technologies make a mix with the telecommunication and information technologies. And this mix makes new revolution. Now in the world of Internet a usual man is likely a writer then a reader.

Internet made a democratic revolution in disseminating the information. But Internet is just the place where the problems of information quality, Copy right and effective ways and means of information retrieval are critical.

In the field of scientific electronic publishing this leads scientific community and publishers to the necessity of the development of new e-publishing models.

2 Changing World of Scientific Publishing and Necessity of New Electronic Publishing Models

The aims of the scientific publishing during centuries were and are:

1. Scientific evaluation of the results of the research.
2. Presentation of the high quality research results & effective scientific methodology to the scientific community. This includes scientific communications and self expression by scientists and scientific schools.
3. Documenting the state of science, logic of science's development and scientific priorities.
4. Support of the high scientific and general level of education.
5. Profit for publisher.

The outcome of this is internal problems for scientific publishing as the institution:

- high prices of scientific edition are not a good way to present the research results to scientific and educational communities,
- the publisher gets profit by selling to the scientists and scientific and educational institutions papers written and evaluated by scientists who are not his employers,
- the members of scientific and educational communities do not wont to pay many to the publishers and arrange there own (without publishers) exchange of scientific papers. And they get approval by the administration of scientific institutions and universities.

Publishing has dramatically changed during last decades. This is due to technological improvements by wide usage of information technologies as the tools for really all the stages of publishing: from the preparation of the text by the author to the presentation the text to the reader. The result is that the text becomes e-text and publishing becomes e-publishing. But the combination of new technologies with old publishing models makes nothing in solving the above internal problems contradictions of scientific publishing. Telecommunications, Internet and WEB make the exchange of e-papers comfortable and cheep for the scientists. For the publishers e-publishing & e-texts open the possibility to sell papers by one to the readers via telecommunications. This is a new model of publishing business, but I believe this new model really is an old one. Really, they operate with the new media at old fashion. The novelty of e-texts is the possibility to use search engines, data mining and texts retrieval technologies to find texts of the interest for the reader and to analyze large corpora of texts as the new readers' services. And sell this services to the user. Actual possibilities of e-texts are much more powerful and exiting than to find the texts with particular words or on the particular subjects. This fantastic power is based on the new information technologies of semiotic and semantic analysis of the texts. The intellectual information technologies are the tools for the texts to tell to the potential reader in the top of the texts' voice. And ones are the base for the new effective models of publishing business with out traditional internal problems.

The electronic editions gives essentially new features to structure and organization for searching information by the reader and the information services providers. Before the computer revolution any edition on a library shelf or under a veil of a dust on a desk, before the reader took it in his hands, meant no more than was written in its catalogue card. (Certainly, we here do not speak about the editions surrounded with light of legends).

Only the electronic edition is capable to speak about itself even in the absence of the reader. The complete dictionary index of the accessible editions, which 30 years back was the dream of any visitor of the scientific library, today has become the present damnation. Let's imagine a reader who wants to find verses about love (about the real love). He will receive a vast list of references on 10, 20, 30,... ways of love, 1001 nights of love, legal, psychological, physiological features of love of sexual minorities, on love to the Fatherland and not love to certain characters. But he searched another matter! His wishes and ideas aspired to something different. He has simply formulated a search image, and the results of the search only hide his idea of love behind a detailed lexical map of the use of the word "love". Fortunately, it is possible to use the skill of the electronic editions to speak into a channel of intelligent, purposeful dialogue with the prospective reader. We shall in this paper discuss the technology ensuring such dialogue on the basis of the automated computed semiotic search and analysis of the textual information.

This dialogue is important not only for the reader, who hungers for the information he wants. It is extremely important for the author or publisher too because of the importance of the authentic prediction of the ways how to understand how the published text is understood by different categories of readers.

3 Intellectual Information Technology for Text Analysis, Search and Retrieval

The non cryptographic e-text as a carrier of information has intention to tell about itself. To use this intention as a base for efficient readers' services one needs appropriate technologies of extraction essential marks of the text from the text as it is. This marks have to be the unique computational (with out human interaction) characterization of the text.

The explosive growth of the information resources available electronically created a necessity for efficient semantic search engines. The traditional methods including keyword search and context search are effectively unable to provide a semantic filtering in sufficiently large data arrays – the resulting search output is still beyond the scope of human analysis. Similar problems are present in an alternative approach – a priori semantic indexing – as it requires compilation and standardizing thesauruses, which poses additional difficulties. Inefficient filtering also puts excessive pressure on the network by tightening the traffic.

To solve the problem of computational characterization of the text we have developed and used the intellectual technologies *Keys to the Text* for semantic search, analysis and indexing of textual information by semiotic analysis of the texts in information resources in natural languages accessible with the help of telecommunications or on electronic carriers of information.

The efficient analysis and search of textual information requires a profound and sophisticated language and text models and essentially new methods. Such models and algorithms were developed by our research team at the Moscow Center of New Information Technologies of the Moscow Medical Academy.

Our search technology is based on a new original two-level model of understanding and interpretation of a text. While the second level requires human interaction and understanding of language semantics, the first semiotic level is the one where purely computational approach offers its help. It turns out that based on combinatorial-statistical analysis, it is possible to synthesize the semiotic pattern of a given text, i.e., to generate a generally small weighted subset of words mostly closely bearing the text's semantics, without referencing to the semantics of the language the text is written in. In particular, the language thesauruses are not used. This phenomenon appears to be true for many European languages (including, for example, Russian and English).

These procedures have been developed by our team and comprise the core of the intellectual technology we offer. In other words, the technology delivers a sufficiently detailed semantic (semiotic) pattern – portrait of a text and can be used for semantic search, classification, and annotation, and development of information resources. While featuring completeness and accuracy of the semantic search and classification, it does not require any form of a priori knowledge of the language besides the language morphology. This facilitates can be applied to virtually any subject area and an extension to more than one language.

The set of tools we developed includes:

- computational semantic indexing, classification, and annotation as a means of search, analysis, and development of information resources in global communication networks and local repositories,
- a possibility for a user to describe the subject area by giving text samples,
- computational structures for characterization, analysis and aggregation the content of large scale collections of the texts.

The technology is oriented to both end-users and information resource providers, developers and publishers.

The offered information technologies of semantic search of the information will present significant interest for various subject domains. Really, our technology of semantic search is subject domain independent, as it does not require the thesauruses and other forms of explanatory dictionaries.

The base for our technology are the algorithms of construction for any text its semiotic ("semantic") pattern – weighted set of words, semiotic mostly strongly connected among themselves in the concrete analyzed text. The word "semantic" is not casually put in inverted commas. When a man makes the analysis and interpretation of this set of words, (received as a result of calculations) their intelligence and connection with subjects, contents and sense of the analyzed text is obvious. But the computations really do not require any semantic information and knowledge of language grammar. These computations use original metrics to extract semiotic connected words in the texts. This metrics (proposed by the author of the paper) needs only combinatorial statistics of the words in the analyzed text and in some set of the texts, representative for language in which the analyzed text is written. The choice of reference set of the texts is equivalent to the formulation of positions of the man, who wants to perceive the concrete text. It is possible to limit such choice to the reference texts of a certain group of carriers of language, for example, professional or political. The problem of forming the reference set of the texts can be treated as the implicit forming of a subset of language adequate to the subject perceiving the text.

The construction of a semantic pattern of the text is based on two basic hypotheses:

1. Semiotic characteristics (semiotic connections of words in the text) determine semantics of the text.
2. To understand or to get a sense of the concrete text it is necessary to determine a reference set of the text, in which context it is necessary to perceive the concrete text.

In essence, it is practically folklore axioms in the linguists, philologists and psychologists societies. It is enough to recollect two classical formulations:

- The man is a style,
- The man is a text.

Validity of the formulated hypotheses proves to be true by high efficiency of the computing analysis of the texts in technology *Keys to the Text.*

Our technology assumes that it is necessary to identify as uniform various forms of each word (for example, one noun in various numbers or a verb in various times, singular or plural number). Such identification enables us to take into account the concrete grammatical forms of the words for construction of semantic patterns of the text. For this purpose the knowledge of the morphology of the language is used. In our technology this procedure (so-called lemmatization) is based on the specific morphological analysis, which allows with high reliability to recognize various forms of concrete words of the given language. Now lemmatization (morphological analysis of the words) is working for Russian and English texts.

The semiotic pattern of the text – the weighted set of words really is unique characteristic of the text. We have looked this in our experiments with Reuters Research Corpus. It includes about 800000 texts of Reuters news. In many thousands of experiments we have conducted with the Corpus we got semantic similarity of two texts more then 96% only for practically identical texts.

4 New Tools, New Possibilities and New Results

The construction of a semantic pattern of the text solves a problem of computational semantic indexing of the textual information adaptively to interests of the user or concrete carrier of the language (professional or political group, individual certain author, edition, group of the editions and so on).

The result of computational indexing is interesting by itself as the means of automatic creation of secondary information resources – lists of key words, which are adequate from the point of view of the concrete reader display the contents and sense of the texts. Simultaneously, the semantic pattern allows allocating in the text mostly important for subjects and contents of the text fragments. That provides automatic generation of the abstracts. Semantic patterns are the base for computational semantic classification of the texts. For this purpose we developed special measure of the semantic affinity based on the above-mentioned metrics of semiotic connections of words in the text.

The solution of the problem of computational construction of a semantic pattern of the text has allowed us to develop new approaches to semantic search and retrieval of textual information. This technology of search and retrieval of information selects just the texts that by the contents and sense really meets the inquiry. The opportunity has appeared to use as inquiry the text selected by user. In technology *Keys to the Text* the semantic pattern of the sample text is computed automatically and adaptively in the interests of the user by the usage of the reference set of the texts, and is treated as an inquiry. The results of search on the inquiry are analyzed and a semantic pattern of each found text is computed. Then the comparison of semantic patterns of the text – sample and found texts – analysis of semantic affinity with the inquiry is carried out. From the results of the analysis the final set of results is formed. The results of the application of technology *Keys to the Text* are highly exact and complete. And it relieves the user from the necessity to solve a very complex problem of describing his/hers interests with just a few words.

The using of semiotic patterns of the texts as unique characteristic of the texts is a good way to solve the problems of the information quality and Copy right. By using *Keys to the Text* technology publisher can find similar texts with out expensive efforts and see how the reader will understand the text two key problems for the publisher.

The development of the tools for generating the semiotic pattern of the text – the unique characteristic of the text is the base for solving the problem of analyzing the large scale corpus of texts and arrangement the specific readers' services.

For the large scale corpus of texts we developed new type of secondary information resources – adaptive interactive thesaurus (AIT). AIT is interactive construction based on the semiotic pattern of the texts of the corpus. The reader begins here work with AIT by any word from the semiotic pattern of any text of the corpus (interactive step). For the reader's usability we join the lists of words in alphabet order and use some other ways to navigate through long list of the words. The second step is done by the computer system. It presents to the reader the list of all the words included in the semiotic pattern of the texts with the word the user have selected (adaptive step). Now reader selects next word of interest. And computer system presents new list of the words included in the semiotic pattern of the texts with two words of reader's selection. We believe this is similar to the thesaurus because the words from the lists are key words for the texts under consideration.

AIT is a tool to solve standard information retrieval problems:

- how to construct the query with not empty set of the search results,
- how to construct the query with reasonable set of the search results,
- how to arrange effective user's feedback for search system and user.

The same time AIT is a tool to develop new services:

- aggregation large scale collections of the texts into secondary resources for efficiently information and essential knowledge retrieval (including terminological and logical structure of the subject domain),
- extracting novelty in the subject domain,
- finding the cross links in different subject domains.

5 New Models for New Possibilities: Old Players – New World

The technologies under consideration in the paper need specific infrastructure for realization. Developed communications, high performance computational resources and users are needed. The possible results will be good suited to the general aims of scientific publishing.

The usage of new information technologies of semiotic analysis of texts is real base for the new models of publishing business. In this models the source of the profit for the publisher is large scale authoritative collections of the texts and not the Copy right for the texts from the collection. Because only large scale collections allow to arrange new services and applications of real value for scientific community. This new World will be based on the new models and will have no internal problems for scientific community and publishing business.