Formation of the Subject Area on the Base of Wikipedia Service

Lande D.V., Andrushchenko V.B., Balagura I.V.
Institute for Information Recording
of National academy of science of Ukraine,
Kyiv, Ukraine
Email: valentyna.andrushchenko@gmail.com

***Abstract*—The paper presents an algorithm for forming the subject domain models based on automatic data analysis of Wikipedia service. It is shown that defined terminology database, dynamically changeable with the network service development, forming a graph structure with nodes, terms, calculated by weight of various terms, concepts for two different criteria - the degree of the nodes, and PageRank. The example shows the adequacy of the proposed approaches, as well as the fact that clusters of terminological networks can be considered as the basis for the identification of individual scientific fields.**

***Keywords*—subject domain models, terminology database, Wikipedia, communication concepts, the network service probing.**

I. Introduction

Information component of contemporary society can be considered one of the most influential and important. Information for today is a product which is absorbed mostly in the Internet. Encyclopedic content resources are also among them, and some of them are claimed in educational, advisory and information purposes. For today one of the most vital tasks is the integration of open and available services and the subsequent formation of new information, visualization and the creation of supporting tools and resources.

We can define the subject domain model as the specially formed network concepts ontology. Creation of a large sector of ontology is a complex scientific and practical problem [1; 2]. The first stage of this process – formation of the terminological foundations of ontology and definition of semantic connections [3]. Subject areas models study, as well as Wikipedia (http://wikipedia.com) service are the subject of many research, confirming the relevance of the investigation [4].

Methods of co-authors networks formation, the definition of significant nodes of the network structure, research citations, as well as relevant case are among them[5].

Authors suggest the method of information networks formation - domain model based automatic monitoring and analysis of reference network information resources. The network is described as the one which meets the terms - headlines of the network encyclopedia Wikipedia.

The purpose of research - the creation of Wikipedia service scunning algorithm for the construction of subject domain models - individual concepts ontologies, which provide the search and testing on individual terms to identify gaps and ways to remove them.

The paper shows description of the theoretical principles, methodology assessment and algorithmic principles of subject areas models creation, including bibliometric field by monitoring and analyzing reference network information resources. To achieve this purpose there was developed a special algorithm of probing Wikipedia service in order to obtain a representative set of terms, concepts like basis (nodes) of the future network.

Due to the implementation of the algorithm it will be possible not only to expand the existing services, but also demonstrate a broad picture of the connection of individual concepts.

Obviously, the network concepts can be large enough if it does not limited by certain theme, corresponding subject area. This feature greatly complicates the perception of the existing network and leads to such effect as issues displacement. To overcome this effect elementary Content Filtering is applied - the only those articles from Wikipedia, which contain basic term defined by the expert is used for analysis. The compliance with these descriptors to determine the size of existing networks – subject domain models and also the dynamics of their formation. In addition, the recognition of clusters in such networks can be considered as the basis for identifying specific scientific areas[1].

II. Research Methodology

Wikipedia was taken for the review, this service is available in the global network and does not presume subscription and also available for download. For the initial access to the system there were applied special terms of target issues for which there are corresponding articles created and edited by authors-experts (Fig. 1).

In a view of these basic terms for a particular subject area there was defined representation of the information in this system. It was also determined that free link switch leads to the effect of the so-called “topic drift”.

Let’s define the “probing of information network” as a sampling of the most important content from big information networks that cannot be scanned by technical reasons.For building of terminological networks it is reasonable to use models that have been tested at peer-to-peer (P2P) networks, which based on equal participants.

In such networks, there are no dedicated servers, and each node (peer) is both: client and server.In many cases, P2P are superimposed (overlays) networks that use existing transport

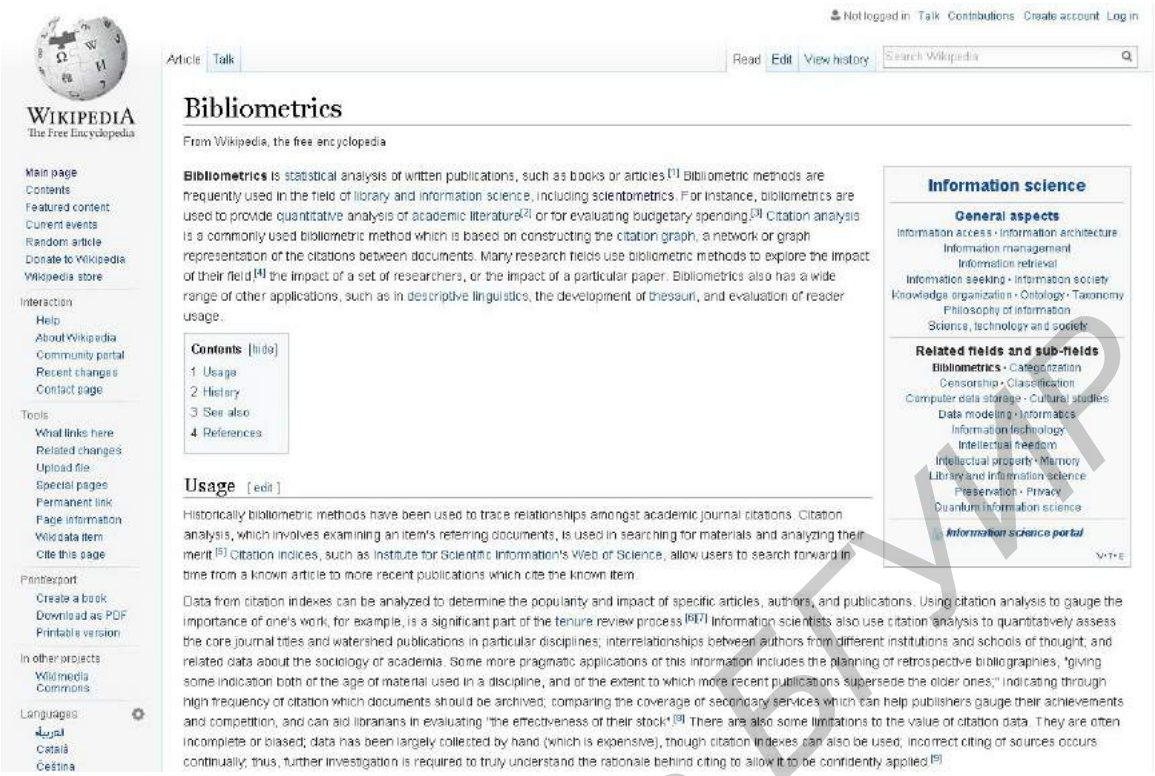

Figure 1. The Wikipedia users interface, article on Bibliometrics

protocols of the Internet.A peer network consist of nodes, each of which interacts only with a subset of other nodes in the network (due to resource constraints).

Several models are used for searching required data in such networks [7]. In the model of Breadth First Search (BFS) the query from a starting node directs to all its neighbors (nearest on certain criteria). When another node receives the request, its local index is looked for and if successful returns a result. Otherwise, the request is transmitted over the network on. In the case of a successful search, a message forms.

- Review (QueryHit) included some information of relevant search sites, and delivered over the network start node.

Another algorithm, the so-called Intelligent Search Mechanism (ISM) provides improved speed and efficiency of information retrievalby taking into account the content of nodes - neighbors to minimize the number of messages between nodes and the number of nodes, the respondents for each search query. In this case, only those components are evaluated for each request, which is the best suited to request. This model is close to the ISM is considered in this work.

The following algorithm was considered for forming models of subject areas according to the Wikipedia service, and involves avoiding of this effect. The algorithm presumes the following steps:

1) On https://www.wikipedia.org/ web page the initial word / phrase is given in the search line;
2) the Search page opens; it provides information about the concept, according to a given search. Initial word / phrase is a vertex of the graph, which will be formed on the results of the probing. All the terms-concepts corresponding the links on the chosen page, are added to the graph. From the initial node the edges-connections are formed;
3) on the current page there are chosen links from the text, without taking into account the information in sections «References», «Notes» and «External Links», as well as not taking into account the «Contents» hyperlink section;
4) all selected words / phrases - hyperlinks are the nodes of the graph;
5) the next transition is executed by the first hyperlink that was defined in the source text;
6) on the page to which the transition has been carried out according to the link, the search of the next word/phrase – the graph vertex is provided;
7) if there is such a word/phrase, the track to the 3d step of the algorithm is provided and respectively from the node-word / phrase of current search new units are being built;
8) if there is no initial word/phrase in the text – the branch is considered to be entirely formed;
9) if during transition to the next word/phrase the jump to the probing page is provided, the word is not added as the graph node, and the reverse connection with the created node is formed.
10) actions on items 3-9 repeat until there will be no initial notion, which is considered as a graph vertex. In such a case the graph is considered to be entirely formed.

According to the described algorithm the process of collect-

ing information from Wikipedia, starting with node-concept stops when according to the algorithm the transition to a new node (with the absence of the basic units) is no longer possible, it means that "cycling" is impossible. The fragment of the program trace which defines the terminological base of the subject domain and corresponding to the above algorithm and basic term Bibliometrics, is shown on Fig. 2.

Bibliometrics

1: Bibliometrics
>!: **Information_science**
>!: **Information_access**
>!: **Information_architecture**
>!: **Information_management**
>!: **Information_retrieval**
>!: **Information_seeking**
>!: **Information_society**
>!: **Knowledge_organization**
>!: **Philosophy_of_information**
>!: **Categorization**
>!: **Information_technology**
>!: **Scientometrics**
>!: **Citation_analysis**
>!: **Citation_graph**

Step: 1; Terms: 14

2: Information_science
>|: Information_theory
>|: Data_science
>|: Library_science
>!: **Informatics**
><: Information_retrieval
><: Information_access
><: Information_architecture
><: Information_management
><: Information_retrieval

Figure 2. Fragment of the program trace

Using the Gephi software there was obtained networks visualization, the networks meet the selected subject areas (Fig. 3).

Obtained characteristics [5] of the formed network Bibliometrics: nodes – 32, connections – 303, density – 0,306.

The most valid two criteria (node degree and PageRank) terms-notions corresponding to the chosen subject area are depicted below in the Table:

The application of cluster analysis can detect most closely related groups of terms that can be used to determine the partial scientific areas by applying a special algorithm that is used in the Gephi system.

III. Conclusion

In the paper there was proposed and implemented an algorithm of subject domains models forming by automatically

Table I. Main terms

Term	Node Degree	PageRank
Information science	18	0.062
Information technology	19	0.061
Information society	19	0.061
Information management	19	0.061
Information retrieval	18	0.055
Information architecture	18	0.055
Information access	17	0.052
Bibliometrics	19	0.051
Categorization	17	0.049
Information seeking	17	0.049
Philosophy of information	17	0.049
Knowledge organization	17	0.049
Informatics	14	0.044
Citation analysis	13	0.041
Citation graph	12	0.036
Information Retrieval	12	0.036
Scientometrics	12	0.036
Information Science	12	0.035
Library science	1	0.008
Information theory	1	0.008
Data science	1	0.008
Computer industry	1	0.008
IT as a service	1	0.008
Manuel Castells	1	0.008
Industrial society	1	0.007
Content management	1	0.007
Knowledge management	1	0.007
Bibliographic coupling	1	0.007
Information systems	1	0.007
Access to Information	1	0.007
Information retrieval applications	1	0.007
Information Architecture Institute	1	0.007

analysis of Wikipedia network service. Such an approach differs from the static models one by consideration of dynamic changes of the resource and considering new concepts, phenomena which appear in bibliometrics in particular. The key elements in such an approach are the names of new articles as the knowledge markers (tags), which are updated by authors – Wikipedia project participants.

It can be mentioned that the Wikipedia system, as the Google Scholar Citations system, which was reviewed earlier [6; 7] is convenient for information access, presumes the creation of users profile for information access, the access is unlimited.

It is necessary to make an accent on the fundamental difference between the proposed automatic terminology models creation from existing networks based on the direct participation of experts in process of selecting specific nodes and links. In the case described in the paper, the researcher uses only crumbs of knowledge represented as the first name, key-term concept to form a network. After that, the program uses the knowledge embedded by (editors) articles authors in Wikipedia, tags defined by internal hyperlinks. In this case, the expert environment is substantially expanded.

The model was applied for the “Bibliometrics” theme in frames of the Wikipedia service, but the suggested approach can be used for other research areas, or for other text arrays, bibliographic databases in particular. Taking into account study and the development of algorithm for the Wikipedia service, the problem of this algorithm applying for other services is set. It is necessary to provide the comparison of the reviewed area in frames of other services for further analysis.

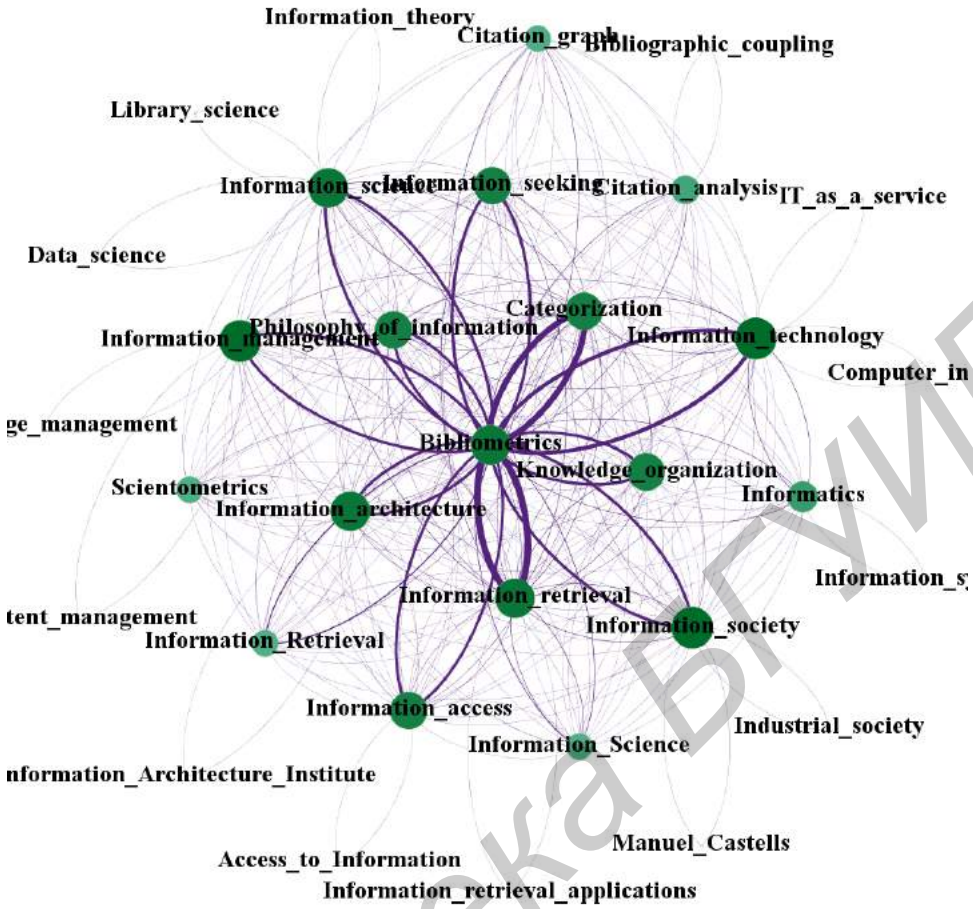

Figure 3. Fragment of the program trace

References

[1] Dobrov, B. V., Solovyov, V. D., Lukashevich, N. V., Ivanov, V. V. (2009), Ontologies and thesauri: models, tools, applications [Ontologii i tezaurusy: modeli, instrumenty, prilozheniia], Binom, Moscow, 173 p.

[2] Lande D.V., snarskii A.A.(2014) Approach to the creation of terminology ontologies // Ontology of designing 2014, N 2(12), – Pp. 83-91.

[3] Chanyshev O.G. (2014) Automatic construction of terminological knowledge base. XVI All-Russian Scientific Conference Digital Libraries: Advanced Methods and Technologies, Dubna, October 13, 2014. – Pp. 85-92.

[4] Zareen Saba Syed, Tim Finin, Anupam Joshi.(2008) Wikipedia as an Ontology for Describing Documents / Proc. 2nd Int. Conf. on Weblogs and Social Media, AAAI Press, March 2008. – Pp. 136-144.

[5] Lande, D. V., Snarskii, A. A., Bezsudnov, I. V. (2014), Internetika: Navigation in complex networks: models and algorithms [Internetika : navigatciia v slozhnykh setiakh : modeli i algoritmy], Librokom (Editorial URSS), Moscow, 264 p.

[6] Lande D.V., Andrushchenko V.B. (2016) Creation of the co-authorship network of law experts on the basis of exploration of the Google Scholar Citations service. Legal informatics, 2016, N 1(46), – Pp. 146-150.

[7] Lande D.V. (2015) Creating a domain model by probing Google Scholar Citations. Ontology of designing, 2015, V. 5, N 3 (17). – Pp. 328-335.

ПОСТРОЕНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ НА БАЗЕ СЕРВИСА WIKIPEDIA

Ланде Д.В., Андрущенко В.Б., Балагура И.В.

Материал статьи содержит описание алгоритма построения предметной области, который основывается на автоматическом анализе информации ресурса Wikipedia. Сервис Wikipedia был выбран для работы, как популярный энциклопедический ресурс открытого доступа. Результатом работы является построение сети понятий, взаимосвязанных между собой, при этом вершиной сети является понятие, задаваемое для исходного поиска. Сесть строится исходя из присутствия заданного слова на тех или иных веб-страницах, переход к которым осуществляется по гиперссылкам со страницы исходного понятия. Реализация алгоритма не только позволит расширить возможности существующего энциклопедического ресурса открытого доступа, но и представить расширенную иллюстрацию взаимосвязей отдельных понятий. В статье приведен пример реализации алгоритма для понятия – Bibliometrics. Также показаны преимущества системы и возможности усовершенствования предложенных приложений.