An Ontology based Visualization Approach for the Joined Interpretation of Bibliometrics and Webometrics Data

Bernd Markscheffel
Ilmenau Institute of Technology, Dept. of Business Informatics
PO Box 100565, 98684 Ilmenau (Germany)
bernd.markscheffel@tu-ilmenau.de

ABSTRACT

Bibliometric analyses enable the measurement of scientific information and allow an evaluation of scientific productivity and efficiency within certain limits. On the other hand an ongoing interest in webometric analysis can be observed. Till now these two parts of informetrics research areas are separated in their interpretation. In this paper we will summarize our experiences in terms of providing a holistic view on both bibliometric and webometric studies with the help of TopicMaps based ontologies. We will explain the problems dealing with the visualization of quantitative aspects of TopicMaps with the help of a special framework. Finally we will give an outlook on the potential of ontologies providing an expanded view on the examined context.

Categories and Subject Descriptors

H5m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Management, Measurement, Documentation.

Keywords

TopicMaps, Bibliometrics, Webometrics, Visualization.

1. INTRODUCTION

The broad range of research areas, the subjects of interests or generally speaking, the facets of informetrics have been discussed in several publications [5, 18, 19, 27]. A common understanding was found in the manner how the research has to be done ("quantitative aspects" [19] "comprising all-metrics studies" [5] or "quantitative studies of…" [18]) and for the subjects of interests, which where summed up by Stock [18] in the categories information user & usage, information itself and information systems. These several pigeonholes are used to describe the several disciplines of informetrics. So, bibliometrics encompasses the measurement of quantitative "properties of documents, and of document related processes" [3] like citation analysis, co-word frequency-analysis or simple counting of publication related properties (e.g. number of publication per author; -per institution; number of (self-)citations; number of authors [20]). It yields at the macro-level (e.g., a whole country) at best general assessments of fields as a whole, for instance, how a country's performance in physics, medicine or immunology is [24]. On the meso-level it breaks down to the research groups, to faculties, chairs or projects. This trend towards disaggregation is not yet completed. But the examination of research groups and chairs is established as a well accepted evaluation unit. And on the micro level bibliometric studies can provide information on productivity of individuals to assist the process of personnel selection in a quite objective way (for example to support tenure-track decisions) or to support managers in the assessment of the research performance of individual scientists [4, 25]. On this level we have the largest number of indicators for providing information for science productivity or related areas.

On the other hand we can observe an ongoing interest in webometric studies, as the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches [2, 21].

The expression power and the number of indicators and also the areas of observation can be seen contrary to the bibliometric levels of observation. Webometric indicators [6] on the micro level are rather less used than on the macro level, because of the lack of interpretation and the minor number of tools, which are supporting micro level analysis. So we can find a larger number of macro level analyses, like university rankings [7, 12, 17], companies or department comparisons [1, 15] than meso- or micro level studies. Hence, bibliometric and webometric studies are actually performed isolated

Main objective of this paper is to describe a framework for the visualization of the federated results of bibliometric and webometric studies, which are considering the same context to provide a solution for a holistic view on the observed area with the help of TopicMaps based ontologies. The informetrics context of our research is described in chapter 2. We summarize our experiences in designing a workflow in chapter 3 - starting from a raw set of bibliometric and webometric data and finishing with the object-oriented structure for the visualization framework. We explain the problems dealing with the visualization of quantitative aspects of TopicMaps with the help of our framework in chapter 4 and finally we give an outlook on the potential of ontologies providing an expanded view on the examined context.

MEDES'11, November 21-24, 2011, San Francisco, USA.
Copyright © 2011 ACM 978-1-4503-1047-5/10/10...$10.00.

2. CONTEXT

We have performed both, a meso-level study concerning the publication behavior of information management chairs in the German speaking world [11] and a webometric study with the same audience and the same spheres of interest to gather information about the external impact, the visibility and popularity of the observed research unit's websites. We have determined as the targets of our investigation the 40 top relevant research units, which are doing their core research in business informatics and information and knowledge management in Germany, Austria and Switzerland.

The time range for the study was 6 years, from 2002 to 2007 and we have chosen Web of Science and Google Scholar as data sources for our investigation. In our analysis we use as the dominant ranking factor the first order h-index (h_1) according to Prathap [14]. The h_1-index allows us to regard the research unit as a (virtual) author and therefore as a whole. But, the h_1-index can be high because the research unit has many researchers that are highly cited, or because the research unit has just a few scientists with a very high h-index [16]. To consider this difference, we use additionally the second order h-index (h_2) as the second ranking factor [14].

The h2-index takes more into account the distribution of the publication behavior in the research unit. Additionally we take into account the total number of citations of the publications of a research unit in the observed field within the observed time range [26] as an extent for the visibility or appreciation of a publication. We use the citation rate as third criteria to further distinguish the research units which have equal h_1- and h_2-indizes. Finally we use the publication rate as an indicator for scientific activities. To perform the analysis of the research units of information management in the German-speaking world we had to evaluate approx. 640 publications and about 1 500 citations with the help of the SCIE and about 2 400 publication with approx. 22 000 citations using Google Scholar.

The target of the observation in the webometrics area was the same as mentioned in the bibliometrics study - the 40 research units in the field of information management. The time range was not important. We performed our webometric study between June and October 2009. Every acquisition of data was repeated five times to reduce the failure due to search engines dynamic changes of the data basis. We used a modification of the well known external Web Impact Factor (WIF), the "research" WIF, which is computed by dividing the number of external Inlinks by the number of (research-) stuff [9] for the ranking of the research units. These data (the results from the bibliometric- and the webometric study) serve as basis for the ontology construction.

3. WORKFLOW

A Part 1: Data Modeling

Based on our experiences in a digital library project [10, 22] we started the integration process with modeling the knowledge structure of the domain. We developed a data model, which supports this transforming process of the raw data into the TopicMap structure on a detailed analysis of the knowledge domain. We have analyzed the data of the bibliometric analysis and generated an Entity Relationship Model (ERM) with the help of this data, to get an idea of the structure and number of relevant topics, which will be later on represented as individual topic nodes in the TopicMap and also of the relationship between them.

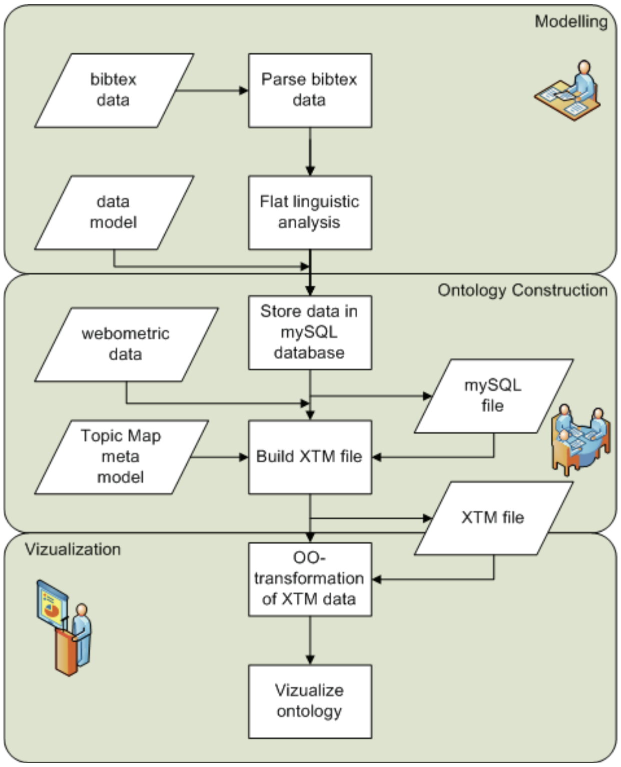

Figure 1. Workflow description from Informetrics data to ontology visualization

This formal representation facilitates the automated generation of a database file for an effective handling of this huge amount of data. On basis of this ERM we have developed a transformation algorithm, which uses the BibTeX files of the bibliometric analysis as standard input. It converts the BibTeX data in a database structure and performs also a flat linguistic analysis using the title of the article to extract the main keywords of the paper and store it also in the database. To get an idea of the bibliometric data a sample cut-out of the BibTeX file of number one ranked research unit is shown in Table 1.

Table 1. Sample cut-out of the BibTeX data

Fachgebiet: Universität Karlsruhe – AIFB / Forschungsgruppe Wissensmanagement, Fachgebietsleiter: Prof. Dr. Rudi Studer – Datenerhebung am: 2009-01-15 Zitationsindex: Google Scholar –
@article{uni-karlsruhe:stsuvo:2002,
AUTHOR = {Studer, Rudi and Sure, York and Volz, Raphael},
TITLE = {Managing User Focused Access to Distributed Knowledge},
JOURNAL = {Journal of Universal Computer Science (JUCS)},
VOLUME = 8,
NUMBER = 6,
YEAR = 2002,
PAGES = {662-672},
Cit = 10,
YearCit = {*},
SelfCit = 5,
AuthCnt = 3}
@article{uni-karlsruhe:maetal:2002a,

B Workflow Part 2: Ontology construction

The next step towards an integrated view is the construction of the TopicMaps-based ontology. The first step was already made with the database creation. The second step is the transformation of the database representation in a valid XTM-file. As lowest common denominator we use research unit around which we will create the ontology. Unfortunately there is no standardized format convention for the storage of webometric results. So it is (actually) necessary to integrate the topics, which are completing the holistic view from the webometric point of view, manually in the TopicMap. Figure 2 shows the role of the ontology layer right in between of the results of our bibliometric and webometric analysis The integration objects from bibliometric point of view are research unit with its several attributes (h_1-index, h_2-index, sum of publications), person (publication rate, citation rate, h-index) and paper (citation, self-citations, number of authors, year). From webometric point of view our ontology integrates the several variations of the Web Impact Factor (WIF_{1-4}).

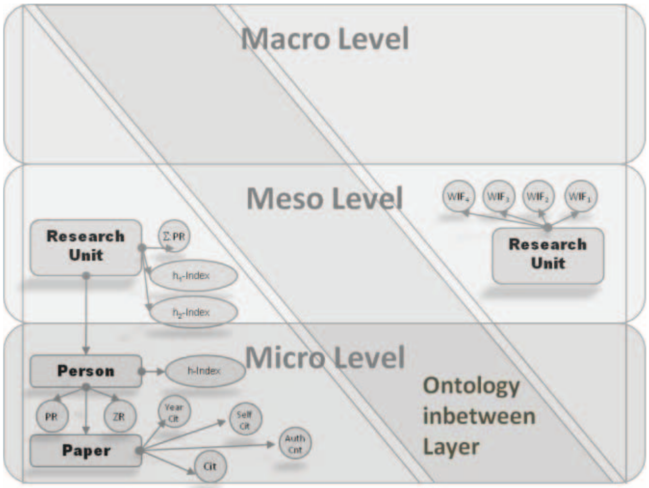

Figure 2. Ontology in between layer

Our ontology is based on the data model which was derived from the bibliometric study and was finally implemented as an XML-TopicMap. TopicMaps is a standard for the representation and interchange of knowledge, with an emphasis on the findability of information [10, 13]. A TopicMap represents information using the following terms:

- Topics, representing any concept, like person, publication, research unit, keyword and index;
- Associations, representing relationships between topics like works_for, published or description, (Associations in TopicMaps are undirected, so one can use Roles for the detailed description of the behavior of a Topic within an association);
- Occurrences representing information resources relevant to a particular topic, e.g. h-index with the scope of 16 for a given author or WIF of 33.48 for a research unit (see Table 2).

As one can see, we have modeled the research unit as a topic with a base name string “University of Karlsruhe“. This topic has several relationships to other topics, and a number of special characteristics concerning the concrete bibliometric and webometric data. To make the later on mentioned problem of visualization a bit easier to solve, we had do find a flexible solution for the modeling of quantitative aspect within the ontology. The sample in Table 2 illustrates our solution. We modeled every (bibliometric/webometric) attribute – value pair as new, separate occurrences. The attributes are defined as scope (see Table 2, e.g. “h-index”, “wif”) and its resourceData (see Table 2, e.g. “16”, “33.48”) as the current value.

Table 2. Sample cut-out of the XML-TopicMap

```
<topic id ="researchunit2">
<instanceOf>
        <topicRef link:href="#researchunit"/>
</instanceOf>
<baseName>
        <baseNameString>University of
        Karlsruhe </baseNameString>
</baseName>
<occurrence>
        <resourceRef   xlink:href=
        "http://www.uni-karlsruhe.de/"/>
</occurrence>
<occurrence>
        <scope><topicRef xlink:href="#indexh-
        index"/></scope>
<resourceData>16</resourceData>
</occurrence>
<occurrence>
        <scope><topicRef
        xlink:href="#indexwif"/></scope>
        <resourceData>33.48</resourceData>
</occurrence>
</topic>
```

C Workflow Part 3: Visualization

A challenging task is to provide an intuitive, easy to use and flexible access to the modeled knowledge of the domain [10, 23]. A graphical visualization of the relevant concepts, their relations and the amount of corresponding subject relevant resources can be a helpful supplement for the illustration of the complex relations within such a holistic view.

To find a solution we have analyzed several visualization tools in the – in our case overlapping – domains of informetrics and TopicMaps. So, we have analyzed the usability and functionality of the following tools:

- two well known bibliometric visualization tools (HistCite[1] and CiteSpace[2]) and of
- three TopicMaps visualizer (Ontopias Omnigator[3], TMNav[4] and our first approach in visualization TMchartis [23]
- and a special solution for visualizing quantitative data (TouchGraph[5]).

[1] http://www.histcite.com
[2] http://clusters.cis.drexel.edu/cchen/citespace/
[3] http://www.ontopia.net/omnigator
[4] http://tm4j.org/tmnav.html
[5] http://touchgraph.com

Main aim of this analysis was to get inspirations and ideas for an improved visualization framework. The results of this analysis can be summarized as follows:

- A TopicMap containing semantic information, tend to be complex and extensive. To support navigation, interpretation and retrieval it is obviously not very helpful to visualize it completely [23].
- A common solution is a subject centered approach, whereby for a selected node all associated concepts are displayed in an automated generated graph. But instead of an automated generated visualization, multiple problem oriented views are needed, which focuses on the individual requirements of the user and the specific problem oriented tasks rather than a generic visualization of the semantic information.
- This approach can help to simplify the interpretation and prevent the user to be overwhelmed by the huge amount of other semantic information. To create such problem-oriented views a human interaction is inevitable. Such an intelligent design approach shifts the focus from the automated generation to the design process where manually visualization information are added, e.g. selecting of important nodes, specification of the node arrangements as well as the highlighting of important aspects.

But the great lack of this approach, and every other TopicMaps-based approach we have analyzed was the absence of any ability for representing quantitative information. This might be due to the concentration on the visualization of semantic relationships and the focus on tree-, network- or graph-oriented-illustrations.

The analysis of the bibliometric tools showed us several possibilities for the visualization of quantitative information. So we found, that the size of the shape or the filling styles of the graphical items are often used for the representation of quantitative information. Moreover, also the arrangement of the topics can be used for a better interpretation of complex information relationship. It would be helpful to store problem-oriented views to support visualization creation process.

With these intentions we have expanded our visualization framework TMchartis [23]. As mentioned before it was developed in the context of the digital library project "DMG-Lib". This new framework toME (topic Map Editor) combines the intelligent design approach with the ability to illustrate quantitative aspects of ontologies. Here are some selected features of this enhanced framework

- free choice of nodes-form (rectangle, triangle, rhombus etc.),
- free choice of filling style (color, pattern, image etc.),
- free choice of line style, dash pattern, size and opacity of a node,
- free choice of line style and dash pattern of associations,
- optionally or permanent display of roles and types of an associations (MouseOver),
- optionally display of information dealing with a specified topic (BaseNames, Occurrences),
- free choice of color scale and node size for topic quantitative indicators,
- Wikipedia and Google interface via browser for better keyword- or other subject illustration.

4. ARCHITECTURE

toME is designed to create multiple problem oriented TopicMaps visualizations. It consists of two major applications developed in Java using TM4J, Hibernate and MySQL. The first part of the application refers to the ontology input: The ontology – which shall be visualized – and which serves as a basis for the several visualization projects is loaded with the help of the TM4J[6] components, which are referring to the according TopicMap-elements

- TM4JNet → TopicMap,
- TM4JNode → Topic,
- TM4JRelation → Association
- TM4jOccurrence → Occurrence.

With the help of Hibernate[7] as persistence provider the ontology is stored in a MySQL-database. For every TopicMap is one database schema available. The second main part of the framework is responsible for the creation of several problem oriented views and the rendering information. It manages the several one or many possible views for one ontology as mySQL-database schema. Such a view is called TMVIEW. In every TMVIEW the rendering information of the ontology elements (Topics, Associations, and Occurrences) are stored as graph of nodes and edges. So, it is possible to create many problem oriented views, depending on the interpretation context. The visualization data are separated stored from the TopicMap data.

The various editor modules are:

- style property manager with its stroking, filling, text_editing, shape_editing, image_editing, and style_template_editing functionality;
- topic element manager with its topic and edge editor and the
- transform tool manager with the transformer, line editor, shape layout creator and aligner, and the
- root tools with selector, zoomer, and positioned functionality.

This functionality is completed with the above explained features for handling of quantitative data. We have added a node scaling tool, a clustering tool and a legend manager which is able to explain the range of values in more detail. Figure 4 illustrates a part of the TopicMap with the three research units (Karlsruhe, Frankfurt and Klagenfurt), authors, a sample of their publications and the corresponding bibliometric and webometric indicators.

[6] TopicMaps for Java (TM4J) is a Java library for processing TopicMaps. TM4J is open source under Apache Foundation license. The goal of the TM4J project is to develop robust tools for creating, manipulating and publishing TopicMaps. It includes a parser, data model, in-memory and persistent storage mechanisms, and a query engine (http://tm4j.org).

[7] Hibernate is a solution for object relational mapping and a persistence management solution. Hibernate maps Java classes to the database tables. It also provides the data query and retrieval facilities. It is an open source project and also a critical component of the JBoss Enterprise Middleware System (JEMS) (http://www.hibernate.org).

5. SUMMARY AND FUTURE WORKS

This paper could be seen as the beginning of a joined interpretation of bibliometric and webometric studies. It allows us the integration of these – in past separated – informetrics fields into a holistic view with the help of TopicMaps based ontologies. But till a professional usage of our visualizing framework a lot of work has to be done to get it out from the prototype state.

A) We had to optimize the workflow, which works on the bibliometric side – with the Bibtex2XTM-conversion – in a good, automated way. The integration of the webometric data in to the TopicMap is actually only possible by hand, which is inacceptable for larger ontologies.

B) The handling of larger data sets – especially the generation of the XTM-file – is a very time consuming process. For a professional usage it is necessary to improve the runtime behavior.

C) Because of the amount of possibilities for the design and composition of the layout it is necessary to support the user via templates which are derived from both from the perceptional point of view and from the informetrics area.

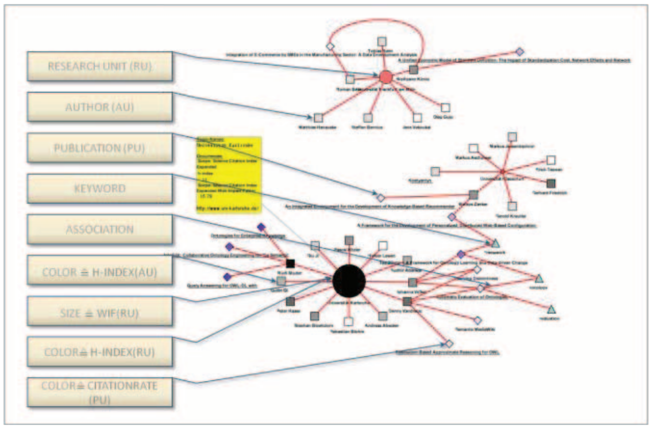

Figure 4. toME-visualizer sample with explanation

Figure 5 shows a possible use case as an example of the capabilities of our approach. One can integrate every topic which is semantically related to one or more topics within the TopicMap. In our example we propose additional

- Alexa Traffic Rank[8] – data for the university, where the research unit is a part of,
- the friend_of_a_friend – data to illustrate the collaboration behavior in more detail for the members of the research unit, and

a Wordnet[9] interface to get an explanation and a context definition of papers keywords.

As mentioned before every semantic related topic can be used with the help of our framework to enhance the interpretation range of bibliometric and/or webometric studies to put the results of these studies in a broader interpretation context and to see the bigger picture.

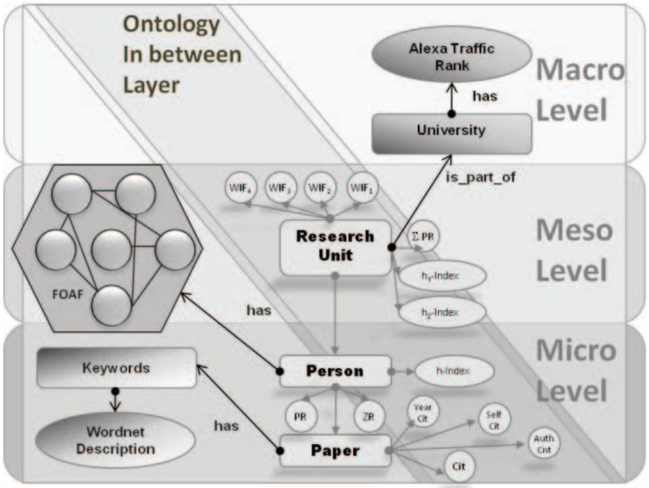

Figure 5. Integration potential of the TopicMaps based ontology

6. ACKNOWLEDGEMENTS

The implementation part of the paper has benefited from insights gained from conversations with Martin Schwarzkopf, who was leading the enhancement and prototyping of toME. Thank you Martin!

7. REFERENCES

[1] Arakaki, M., and Willett, P. 2009. Webometric analysis of departments of librarianship and information science: a follow-up study. Journal of Information Science, 2 (2009), 143-152.

[2] Björneborn, L., and Ingwersen, P. 2004. Towards a basic framework for webometrics. Journal of American Society for Information Science and Technology, 14 (2004), 1216–1227.

[3] Borgman, C. L., and Furner, J. 2004. Scholarly communication and bibliometrics. Annual Review of Information Science and Technology, 36 (2004), 3–72.

[4] Costas, R., and Bordons, M. 2004. Bibliometric indicators at the micro-level: some results in the area of natural resources at the Spanish CSIC, Research Evaluation, 2 (2005), 110–120.

[5] Egghe, L. 2005. Expansion of the field of informetrics: Origins and consequences. Information Processing & Management, 41 (2005), 1311–1316.

[6] Ingwersen, P. 1998. The calculation of Web impact factors. Journal of Documentation, 2 (1998), 236–243.

[7] Kumar, J. S., Chandra, B. S., and Parthasarathi, M. 2010 Web-based ranking and link analysis of Central Universities in India: A webometric analysis. Information Studies, 1 (2010).

[8] Li, X. 2010. A review of the development and application of the web impact factor. Online Information Review 27, 6 (2010), 407–417.

[9] Li, X., Thelwall, M., Musgrove, P., and Wilkinson, D. 2003. The relationship between the links/Web Impact Factors of computer science departments in UK and their RAE

[8] http://www.alexa.com/

[9] http://wordnet.princeton.edu/

(Research Assessment Exercise) ranking in 2001. Scientometrics 57, 2 (2003), 239–255.

[10] Markscheffel, B. Thomas, H., and Redmann, T. 2009. Developing Topic Maps Applications: Lessons Learned from a Digital Library Project. In Kommers, P., and Isaías, P. (Eds.). Proceedings of the IADIS International Conference e-Society 2009. Barcelona, Spain, Feb. 25-28, (2009), 51–59.

[11] Markscheffel, B., and Giese, S. 2009. The use of meso level bibliometric analysis in research performance assessment and monitoring of information management chairs in Germany, Austria and Switzerland. Hou, H., Wang, B., Liu, S., Hu, Z., Zhang, X., and Li, M. (Eds.). Proceedings of WIS 2009 and Fifth International Conference on Webometrics, Informetrics and Scientometrics & Tenth COLLNET Meeting, Dalian, 2009. September 13–16, Dalian, China.

[12] Ortega, J. L., and Aguillo, I. F. 2009. Mapping world-class universities on the web. Information Processing & Management 2, March (2009), 272–279.

[13] Park, J., and Hunting, S. 2002. XML Topic Maps: Creating and Using Topic Maps for the Web. Addison-Wesley Longman Publishing Co.

[14] Prathap, G. 2006. Hirsch-type indices for ranking institutions' scientific research output. Current Science, 91, 11 (2006), 1439.

[15] Qui, J. P., Chen, J. Q., and Duan, Y. F. 2009. An Analysis of E-Commerce of Travel Agency Websites in China - Link Analysis and Web Impact Factors. Journal of Guang-dong Radio & Television University 2 (2009).

[16] Rosseau, R. 2008. Reflections on recent developments of the h-index. In: Kretschmer, H. and Havemann, F. (Eds.). Proceedings of WIS 2008, Berlin. Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting. Berlin, 2008, 1–8.

[17] Smith, A., & Thelwall, M. (2002). Web impact factors for Australasian universities. Scientometrics, 3, (2002), 363–380.

[18] Stock, W., and Weber, S. 2006. Facets of informetrics. Information - Wissenschaft und Praxis, 8 (2006), 385–389.

[19] Tague-Sutcliffe, J. 1992. An introduction to informetrics. Information Processing & Management, 28 (1992), 1–3.

[20] Thelwall, M. 2008. Bibliometrics to webometrics. Journal of Information Science, 4, (2008), 605–621.

[21] Thelwall, M., Vaughan, L., and Björneborn, L. 2005. Webometrics. Annual Review of Information Science and Technology, 39 (2005), 81–135.

[22] Thomas, H., Brecht, R., Markscheffel, B., Redmann, T., Bode, S., and Spekowius, K. 2007. TMchartis - a Toolset for Designing Multiple Visualizations for Topic Maps. Proceedings of 3th International Conference on Topic Maps Research and Applications - Scaling Topic Maps (TMRA), Leipzig, Germany, (2007).

[23] Thomas, H., Redmann, T., and Markscheffel, B. 2007. Controlled semantic tagging? How can Topic Maps support subject indexing in digital libraries? Shoniregun, C A. and Logvynovskiy, A. (Eds.), In: Proceedings of the International Conference on Information Society (i-Society), (2007), 346–352.

[24] van Raan, F. J. 2003. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. Technikfolgenabschätzung, Theorie und Praxis/Technology Assessment, Theory and Practice, 12 (2003), 20–29.

[25] Vinkler, P. 1988. An Attempt of surveying and classifying bibliometric indicators for scientometric purposes. Scientometrics 13, 5-6 (1988), 239–259.

[26] Vinkler, P. 1995. Some aspects of the evaluation of scientific and related performances of individuals. Scientometrics 32, 2 (1995), 109–116.

[27] Wilson, C S. 1999. Informetrics. Annual Review of Information Science and Technology 34, (1999), 107–247.