Articles

Intelligent Tagging and Search as a Fully Automated System

Maria Marzovanova*

Summary:

Taxonomic structures and folksonomy tags are well-known mechanisms used in systems for organizing unstructured content. But they come from different worlds, taxonomies tend to be highly professional, both referring to their building process and their usage, while folksonomies are more user-oriented and therefore they suffer from the lack of professionalism. A new structure of terms called metataxonomy has been developed and the advantages of its implementation has been already demonstrated. This paper briefly presents the idea laying behind the concept for intelligent tagging and search based on this extended taxonomic structure in order to depict the missing part that will bring that system to the state of being thorough.

Key words: tagging, search, metataxonomy, text documents

JEL Classification: C61, C8, C88

1. Introduction

The need for "structuring unstructured data" and the problems that the researchers in the area have confronted and still confront are not news to us. Therefore there is a wide variety of researches and projects, as well as developed and working tools and systems that should be considered as those laying the laws and principles in knowledge organization.

Knowledge organization systems are intended to help search engines by applying diverse content-arranging techniques, most of them using terms in some form as an organizing unit. Depending on the base structure of the terms, the knowledge organization systems can be categorized as:

- Taxonomy-based – they use a taxonomy structure developed by an expert on the subject which are difficult to customize by the average user,
- Folksonomy-based – they use user generated term collection, which has no structure but as it is highly personalized users find it easy and likable,
- Combinations – they use taxonomies, which are result of some transformation of folksonomies (Christant, 2010) (Kiu & E.Tsui, 2010).

Irrespective of the system's category, they all have something in common, and this is the organizing principle. There is a collection of terms, whether structured or not, and a process of content categorization. In different systems that process consists of different number of steps but at the end there is one idea standing behind it - the unstructured content is either distributed or associated with one or more of the keywords in the collection.

Intelligent tagging and search is a set of

* Assistant, PhD, Department Information Technologies and Communications, University of National and World Economy, email: mmarzovanova@unwe.bg

mechanism that introduces a new approach to solving the problems with structuring unstructured data. The main features which make the concept of intelligent tagging and search different can be found in improving the structure of terms, in automating the process of categorization of the unstructured content, and especially in the way the system actually uses the relations within the structure.

The purpose of this article is to present the picture in which intelligent tagging and search is situated. This comprehension along with a summary of the main principles laying behind the development of a system for intelligent tagging and search will depict what is missing in order to complete the automated cycle of creating meta-taxonomies and using them for different search and text analysis purposes in the context of a system of that kind. Considering the high speed at which that kind of content is growing we need systems for organizing it in a way that is

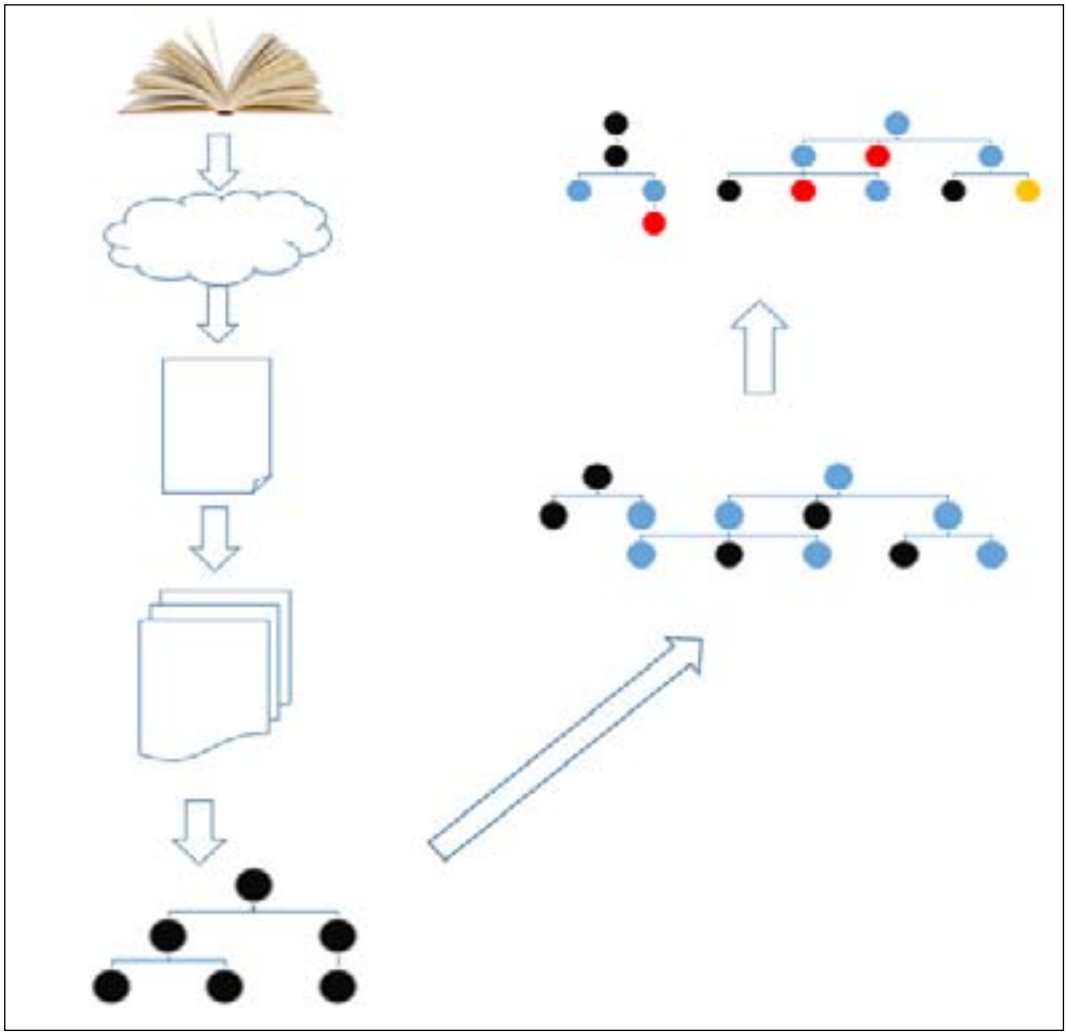

Fig. 1. Metataxonomy levels of evolution

fast and self-developing with minimum, or better, no human intervention.

2. Metataxonomy

Metataxonomy is the structure of terms used for intelligent tagging and search. It was firstly introduced with the name "agile taxonomy" and a description of its structure is given in a paper called "Agile taxonomy/folksonomy of economics in Bulgarian language" (Marzovanova, 2012). After refining the idea, it was later called a metataxonomy in the dissertation of the author. To catch the essence of the metataxonomy we should say that this is a structure developed as a combination of taxonomy and folksonomy features, but having a detailed look in the process of its development, it's not a transformation of a taxonomy into a folksonomy, it is an extension. The figure below illustrates the different stages of creating a metataxonomy starting from a book. Each and every stage changes the "shape" and the number of terms involved in the result, which allows the author to call these stages – levels of evolution. The focus is not on refining the initial structure, it's on extending and expanding it in order to make difference in the information that it holds. Starting with a basic taxonomy of terms, developed according to one trusted source, gives a feeling of control and reliability. On the next level the terms included in the vocabulary are connected with their translations in English. That ordered collection is then enriched with a new kind of relation – synonyms, selected from trusted dictionaries over the web. On the next level comes the user of the taxonomy, who can reorder and add new terminology if necessary with a logical, synonym connection between the new terms or even his personal type. On this final stage comes the freedom for the user, a feature considered as the main advantage of the folksonomies.

There is not only a wide range of words that can be included but the more important thing is that there is a wide range of relationships that put a meaning to position a certain term among the others. These relations make the structure so different and allow the system using it to implement more and more functionality related to knowledge organization and even to text analysis.

3. Metataxonomy-based systems

Intelligent tagging and search opened a new category in the classification mentioned above, metataxonomy-based systems. There are some specifics that make this category an independent one – the structure of terms, called metataxonomy, the degree of automation achieved and the ability of a system to read a text before deciding how to categorize it. All the main points involved in the concept of intelligent tagging and search are outlined in the picture below.

A system from the metataxonomy-based category should provide instruments for creating and storing metataxonomies. A database, specifically designed for storing terms and their relations, making it possible to execute queries heading from the bottom of the classification to the top and reverse. Another important feature is the common functions such as copy, paste, move and delete nodes.

The rest of the required functionality includes the system's ability to work with unstructured data (Marzovanova, et al., 2012) (Marzovanova, 2013). This actually substitutes reading the content of a document, finding keywords and making a decision in which category to locate it. Main features here are:

- Access to the content of diverse file formats
- Creating annotations and applying them on the content
- Creating indexes for each pair annotation – annotated text

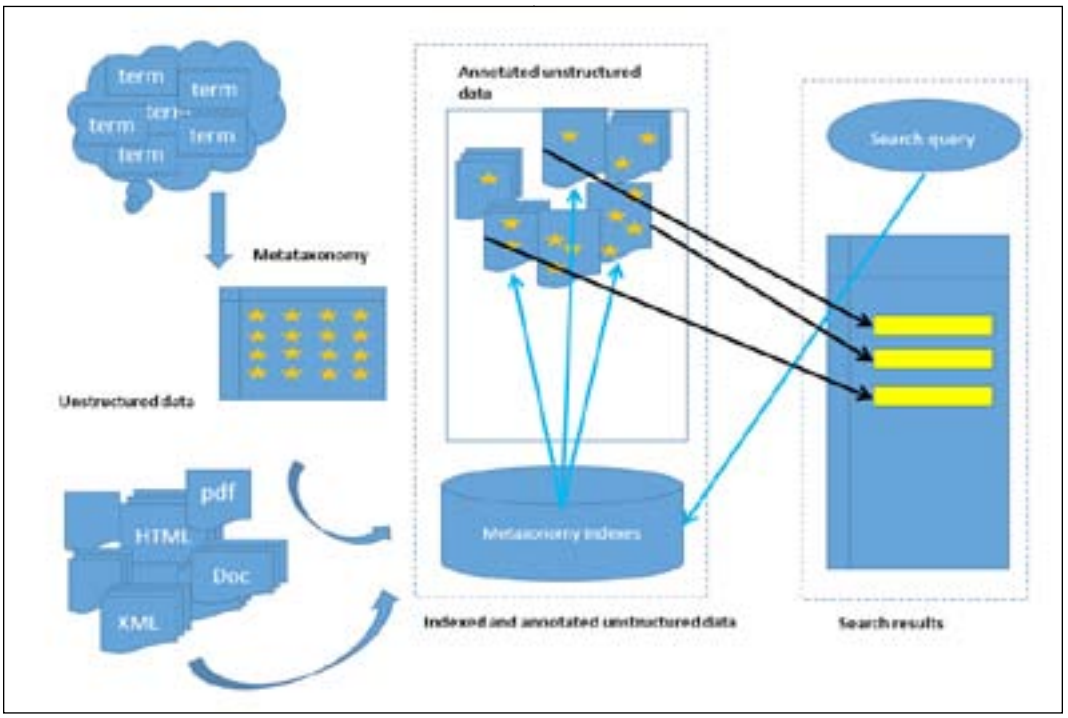

Fig. 2. The concept of intelligent tagging and search

- Search based on both the indexes and the logic behind the relations between the terms, part of that indexes

What is important is how all these specifics mentioned about the intelligent tagging and search appear in the whole picture of knowledge organization. When an unordered collection of terms becomes a metataxonomy it adds meaning to every single word, a meaning which is defined in a concrete context. More specifically, that context has two main aspects:

- Context that is derived from the connections (logical, synonym, translation, hierarchical). Based on the type of the relationship we can conclude if two terms are equal or one of them includes the other, for a start.
- Context derived from the other terms on the same level in the same node. There are plenty of examples for words that spell the same, but have very different meaning. Well, without reading the surrounding text, one cannot be sure which meaning is intended in the specific case.

So having in hand a wide collection of words, arranged obeying certain rules and restrictions, the metataxonomy can provide a precise logic for defining a term's meaning based exactly on those rules.

4. The missing segment

While the concept of intelligent tagging and search was evolving, we developed a prototype system. The automated process of creating a metataxonomy there needs a traditional taxonomy as a start, this taxonomy was created manually.

The outlining so far, implies that for the proper functioning of such a system, there should be a deliberative taxonomy to start the process on. Narrowing the human intervention in building the initial taxonomy is one step closer to the aim. There are

a lot of rules that should be followed when speaking of building a considerate taxonomy. The process of manual taxonomy development is time consuming as much as prone to mistakes if done with less than the expected attention and devotion. First of all, a taxonomic structure needs a wide range of words. Then the collection of terms needs to be refined carefully not to include elements which are outside the domain of the subject being studied or are so common in the literature that they have lost their meaning and information value. Once the vocabulary is clear, it should be organized as a hierarchy depending on relationships between the words and analyzing the strength of that connection. Finally, the resultant taxonomy is reviewed and refined multiple times until it reaches the expected quality.

So here is the name of the missing segment - automated taxonomy generation. There are some researches on creating taxonomies for different purpose. In 2002 Christian Blaschke publishes a method, which generates classifications of gene-product functions using bibliographic information (Blaschke, C., Valencia, A., 2002). Then in 2003 on the topic works Raghuram Krishnapuram, who talks about the problems and the potential of automating the process of taxonomy building (Krishnapuram, R., Kimmamuru K, 2003). More specifically his article focuses on several approaches for taxonomy generation and presents a detailed view of the problems involved. 2004 comes with Sanchez (Sanchez, D., Moreno, A., 2004) who presents a methodology for information retrieval from the internet and creating a taxonomy of terms and web resources for a certain domain. Later in 2007 a new idea becomes public, Schwarzkopf proposes an approach for using data from social tagging systems, such as del.icio.us, as foundation of user adaptation and then the taxonomy to be extracted out of the tag cloud (Schwarzkopf, 2007). They will then use the personalized taxonomy giving the user his own tag collection.

To conclude, the idea to automate the taxonomy creating process is not new and it is possible. Following the ideas presented in the past we can define some main steps constituting the creation of a hierarchy of terms out of a text.

5. Aspects for future study: Creating taxonomies from text

Different algorithms for automated creating of taxonomies have been published (Camiña, 2009) (Chuang S., Chien L., 2002) (Henschel A., Woon W., Wachter, T., Madnick, S., 2009) (Heymann, P., Garcia-Molina, H.,, 2006), an overview of the literature outlined the process in its common appearance.

There are two groups of steps – data preparation and applying an algorithm for taxonomy generation. Data preparation starts with terminology extraction. One way of fulfilling this aim is calculating the frequency of words appearing in a text and the frequency of some words appearing near one another. As a result we have numbers corresponding to each word (Matsou, Y., 2003)

The next step is similarity evaluation, it begin with building a graph including the terms that accomplished the frequency analysis. The similarity is a metric that can be estimated using different approaches, but most appropriate in terms of the current discussion are cosine similarity and Google Similarity Distance (aNGD and sNGD) (Cilibrasi R.L., Vitanyi P. M. B., 2007).

The graph representation transforms into a matrix of distances. It is always a square matrix which contains the calculated distances between each pair of words. Once the distance matrix is created it is time to apply a taxonomy generation algorithm.

Most of the algorithms create a tree out of the information constituting the matrix, but for this they need a starting node (term) to be selected. It can be defined manually or using a mechanism for finding the most central node either by calculating Betweenness centrality or Closeness centrality.

Algorithms for taxonomy generation are also in diverse variety and they should be thoroughly studied in order to define their advantages and evaluate which one is most appropriate for the needs of the system for intelligent tagging and search.

Conclusion

Intelligent tagging and search is a solution to a current problem, structuring unstructured data, data that multiplies every second and becomes useless in the next only because it's hard to be processed and analyzed with that speed.

Implementing automated taxonomy creation in the system for intelligent tagging and search will not only save time and reduce efforts, but it opens a new field for development. With the ability to read a text and arrange new term nodes with words and relationships, the system becomes self-learning and open to consume new ideas and techniques.

References

Blaschke, C., Valencia, A., 2002. Automatic Ontology Construction from the Literature. *Genome Informatics, Volume 13,*. s.l.:s.n.

Camiña, S. L., 2009. A Comparison of Taxonomy Generation Techniques Using Bibliometric Methods: Applied to Research Strategy Formulation. s.l.:s.n.

Christant, F., 2010. Tagonomy - Taxonomy meets Folksonomy. [Online] Available at: http://www.ferdychristant.com/blog//articles/DOMM-86CCFU

Chuang S., Chien L., 2002. Towards Automatic Generation of Query Taxonomy: A Hierarchical Query Clustering Approach. s.l.:s.n.

Cilibrasi R.L., Vitanyi P. M. B., 2007. The google similarity distance. *IEEE Trans. Knowledge and Data Engineering.* s.l.:s.n.

Henschel A., Woon W., Wachter, T., Madnick, S., 2009. Comparison of Generality Based Algorithm Variants for Automatic Taxonomy Generation. s.l.:s.n.

Heymann, P., Garcia-Molina, H.,, 2006. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems.. s.l.:s.n.

Kiu, C.-C. & E.Tsui, 2010. TaxoFolk: A hybrid taxonomy - folksonomy classification for enhanced knowledge navigation. *Knowledge Management Research & Practice,* Volume 8, pp. 24-32.

Krishnapuram, R., Kimmamuru K, 2003. Automatic Taxonomy Generation: Issues and Possiblities. *Lecture Notes in Computer Science.* Berlin: s.n.

Marzovanova, M., 2012. Agile Taxonomy/ Folksonomy of Economics in Bulgarian Language. Sofia, s.n.

Marzovanova, M., 2013. Advantages of using a system for intelligent tagging and search in unstructured data. Sofia, s.n.

Marzovanova, M., 2013. Creating annotations and indexes in System for intelligent tagging and search in unstructured *data.* s.l., s.n.

Marzovanova, M., Tsoykova, T. & Kisimov, P. D. V., 2012. *INFORMATION SYSTEM FOR TAGGING, INDEXING AND SEARCH OF UNSTRUCTURED DATA.* Sofia, s.n.

Matsou, Y., 2003. Keyword Extraction from a Single Document. s.l.:s.n.

Sanchez, D., Moreno, A., 2004. Automatic Generation of Taxonomies, from the *WWW. Practical Aspects of Knowledge Management.* s.l.:s.n.

Schwarzkopf, E., 2007. Mining the Structure of Tag Spaces for User Modeling. Data Mining for User Modeling, International Conference on User Modeling. s.l.:s.n.