Facet analysis as conceptual modeling of hypertexts: methodological proposal for the management of semantic content in digital libraries

Gercina Angela Borém Oliveira Lima[1] and Benildes Coura Moreira dos Santos Maculan[2]

[1] PHD in Information Science, Professor of the Information Science College (UFMG)
[2] PHD student of Information Science, Professor of the Information Science College (UFMG)

Abstract: This study aims at the construction of a semantically structured conceptual model to help the organization and representation of human knowledge in hypertextual systems, based on the Theory of Facet Analysis and Conceptual Map. A second step in this study is the application of the semantic model to create a prototype called Hypertext Map (Mapa Hipertextual - MHTX) which will be used to implement the BTDECI – UFMG (Thesis and Dissertation Library of UFMG´s School of Information Science). Facet analysis was used to categorize the index terms, and established the relationship among them expressed by the links. It was chosen a single knowledge unit, namely a doctoral thesis on the own area of subject analysis (NAVES, 2000) to construct a conceptual model prior to the technological implementation of its prototype..
Keywords: Faceted analysis, Conceptual modeling, Hypertext, Academic documents

1. Introduction

Since the late 1980s, several researchers have started to study the possibility of using the theories of classification – mainly the theory of facet analysis, created by the Indian Ranganathan – in the conceptual organization of systems of hypertext. Like the faceted classification, systems of hypertext also aim at the organization of knowledge through the organization of the concepts and the relations among them, which makes possible the mapping of an area of subject and the inclusion of new concepts, without implying a structural change of the system. A lot of authors mention the faceted classification as a tool which can help in the representation of the intellectual content in systems of hypertexts (Duncan (1989); Ingwersen and Wormell (1992); Kwasnik (1992); Ellis (1996);

Received: 21.5.2014 / Accepted: 1.12.2014 ISSN 2241-1925

Star (1996); Santos (1996); Pollitt (1997); Glassel (1998); Priss and Jacob (1999); Ellis and Vasconcelos (1999); Koehler (2001); Campos (2001)).

The conceptual modeling is an important phase in information organization for systems of hypertext. Facet analysis recognizes several aspects in just one subject, and tries to synthesize these aspects in such a way that it can describe them in the most appropriate way. This fact is relevant for the non-linear approach to hypertext because it enables the user to see a certain subject from different perspectives, making possible the reunification of a similar knowledge as part of the whole instead of being subordinate within a hierarchy.

The use of facet analysis served as groundwork for the creation of an interactive tool, with an organizational power of semantic content in complete documents in bases of hypertextual data and with a possibility of a retrieval in effective context.

The digital prototype, called Hypertextual Model for Document Organization (MHTX), consists of a semantic map called Conceptual Map (CM), and of the Expanded Summary (ES), an instrument formed by the summary of the thesis, to which access points were aggregated. The hypertext model was installed in digital format in a database which holds the selected thesis to be object of this research, and after which will also hold other theses and dissertations as complete scanned texts, which belong to the Graduate Program in Information Science of UFMG. In the next topic, we describe briefly the theory of facet analysis and the construction of the conceptual structure, bearing in mind the Hypertextual Map (MHTX) as result of the PHD thesis of this article's author.

2. Utilization of facet analysis in conceptual modeling for hypertext management of documents

Shiyali Ramamrita Ranganathan (1892-1972), an Indian mathematician who became a librarian, was one of the scholars who contributed greatly to library science theories in the XX century, especially in the field of subject classification. His Colon Classification system, published in 1933, emerged from his dissatisfaction with the Dewey's Decimal Classification and Universal Decimal Classification systems. After certifying, in five different periodicals, that most of the subjects discussed were compound subjects, Ranganathan formulated his Colon Classification, also known as Faceted Classification or Analytic-Synthetic Classification. He created a nearly mathematical system, organizing knowledge in such a way that the compound subjects, synthetically, emerged from "elementary" concepts. Ranganathan published six editions of this system, and died in 1972, when the seventh edition was to be published. In addition to this classification system, Ranganathan published *Elements of Library Classifications* (1945; 3 ed. 1962) and *Prolegomena to Library Classification* (1967), works which are considered referential by scholars of classification. Among the principles introduced by Ranganathan, the most well-known is the principle of facet analysis (subdivisions of subjects into their component parts), and synthesis (reconfiguration of these parts to classify a document appropriately). In his Faceted Classification, Ranganathan identified

five categories: Personality (P), Material (M), Energy (E), Space (S), and Time (T), mnemonically known as PMEST. This order of mention is based on the idea of decreasing concretivity. Although the Colon Classification system has not been accepted worldwide, the theory of facet analysis and synthesis proposed by Ranganathan has become an important theoretical basis for the area of subject analysis in the XX century (Ranganathan, 1985, p. 86).

The faceted classification comprises principles and techniques for organization systems and information retrieval. A faceted system recognizes many aspects in just one subject, and tries to synthesize these aspects in order to describe them in a more appropriate way. Ranganathan showed that the relation among these subjects, made through the symbol of the two points, can be boundless, proving that knowledge can be multidimensional and that these relations can take different directions, depending on the synthesis among various multiple concepts (Vickery, 1980, p. 209).

In order to develop a faceted classification, one examines the literature of the subject to identify its concepts and terms, establishing their features and facets. After collecting and defining the terminology of the subject, the terms are analysed and distributed in facets. As it has already been defined, a facet is the collection of terms that have the same relationship with the global subject, reflecting the application of a basic principle of division. The facets obtained are inherent within the subject; and, within each facet, the terms which are part of them are susceptible to new groupings by the application of other divisional characteristics, generating the subfacets. The terms, in the subfacets, are mutually exclusive, i.e., they must not overlap one another in the formation of compound subjects. When the facets and subfacets are established, it is important to determine the order in which they are presented in the classification system. Then, we sort all the elements in order of filing, which enables one to place the general subject before the specific one. After these phases, the system is ready to receive a notation, which must be flexible to make possible the inclusion of new classes. Finally, we compile an index with all the terms and their respective notations (Piedade, 1983, p. 80, Barbosa,1972, p.76).

In order to check the applicability of the facet analysis in conceptual modeling for hypertext management of documents, we have chosen a doctoral thesis of the Graduate Program in Information Science (Information Science College of UFMG). The choice of the thesis as sampling for the conversion of a linear text into a non linear one occurred according to the following criteria: (1) the fact that theses and dissertations are written linearly; (2) the difficulty of the content author to work directly and simultaneously with the technology author; (3) the impossibility of writing theses in both linear and non-linear ways; and, finally, (4) the possibility of application of this model in the BTDECI – UFMG (Biblioteca de Teses e Dissertações do Programa de Pós-Graduação em Ciência da Informação da UFMG – Thesis and Dissertation Library of the Graduate Program in Information Science of UFMG).

3. The application of facet analysis in the conceptual structure of the hypertextual prototype map – MHTX

The methodological procedures for the conceptual implementation through the theory of facet analyses included, sequentially, the following phases:

1. identification of the basic work document (the thesis);
2. reading of the thesis;
3. facet analysis of the subject of the thesis: the selection of its relevant terms and categories (facets); the recognition of subfacets; the ordination of facets, subfacets and focuses to be presented in the conceptual map; and, finally, the organization of all the terms and their relations;
4. creation of the Conceptual Map (CM) with its links and relations.

For the faceting process of the thesis chosen for this study, we used the technique of facet analysis, based on two processes: (a) analysis, which occurs from the identification of the relevant concepts; (b) classification of concepts in categories, in which each category represents a characteristic. It is important to point out that the process which refers to the synthesis, which occurs when each concept that belongs to these categories matches with another one in order to express a compound subject, has not been implemented in the present thesis.

For the formation of categories, we used the normative principles of the field of ideas which are present in the work *Simplified Model for Facet Analysis: Ranghanathan 101*, compiled by Spiteri (1998), in which the author discussed and synthesized principles established previously on two research fronts: those initially presented by S. R. Ranganathan, and those presented afterwards by the *Classification Research Group – CRG* in London (1952).

In the field of ideas, several procedures occur: (1) the process of subject definition; (2) selection of characteristics which constitute the subject; (3) selection of a model for the mapping of information about the concepts; (4) grouping and division of these concepts according to their common and different characteristics; and, (5) organization and arrangement of groups and subgroups. The principles which rule these procedures have been used to ground the faceting process, aiming to guarantee the determination of the content of bordering bases in relation to subjects which interest the users, the appropriate selection of standards for subject categorization, and, finally, the organization of these categories into an appropriate sequence.

FAT, Facet Analysis Theory, is a model with a deductive methodological nature, in which, in the first place, the domain/context is considered in order to, afterwards, select the representative terms of an area. Thus, it has mechanisms of representation for working at conceptual levels in the formation of categories, from which the concepts are sorted in order to form classes of concepts.

In the analysis of the faceting process of the thesis subject, we made a detailed mapping of the document for correct identification of its theme, and made the conceptual analysis to determine its basic and isolated subjects. For the indexing, we considered the text as a whole, having all the terms that constitute the document in the author's language, since they are considered relevant. Because it is a conceptual modeling, we have disregarded all the

collective entities and names of authors mentioned in the thesis. In the process of analyzing the thesis terminology, we took into consideration the notion of category for analysis and classification of document subjects and for the organization of these subjects (isolated) in a conceptual structure. For the formation of subject (analysis), we used the *Prolegomena to Library Classification* (Ranganathan, 1967, Part P). For the formation of categories (synthesis), we used Spiteri's Simplified Model (1998).

For the formation of subjects, Ranganathan (1967, p.351) proposes five methods: **dissection**, **lamination**, **denudation**, **reunion/aggregation** and **superposition**, which have been recently discussed by Campos (2000, p. 68). In the construction of the conceptual model of this thesis, only two methods were used to form the subjects: **dissection** and **denudation.** The other three methods were used because they deal with the formation of compound subjects (**lamination**), formation of complex subjects (**reunion/aggregation**) and the superposition of two isolated ideas (**superposition**) not related to the object. We have not used these last three types of subject formation because we have used the natural language of the document. In other words, the terms selected remained in the original language of the author, and no controlled vocabulary for the standardization of the terminology was used.

The method of **dissection** has been used to ascertain that a unique domain encompasses the thesis and its parts. From this procedure, we came to the conclusion that the universe of the basic subject of the thesis is "knowledge organization".

In the method of **denudation**, the progressive decrease of extension occurs, as well as the increase of the depth of the basic subject or of an isolated idea. The successive result of **denudation** enables the formation of chains which start to represent the specific core of a basic subject or of an isolated idea.

In the application of the **dissection** and **denudation** methods for subject formation suggested by Ranganathan, some steps were followed and some rules were set, in order to help the standardization of the mapping of concepts in the faceted modeling:

a. detailed reading of all parts of the document;
b. identification of concepts which better represent the semantic content of the document;
c. selection of the most relevant terms in the document for the knowledge organization;
d. selection of terms with a meaningful semantic content which has a bigger amount of textual information;
e. selection of terms which, although they do not have a large amount of textual information, are relevant in the construction of the conceptual modeling to represent the subject;
f. representation of selected concepts, such as they appear in the natural language of the text, without the necessity of a translation that uses controlled vocabulary.

In a second step, in the process of category formation, we proceeded to the assessment and the definition of each concept, with analysis and definition of the isolated terms according to their common characteristics. According to Raghavan (1985, p.27), the analysis of concepts and the relation among them found in a specific subject serve as ground for a structure which expresses the meaning and the relation among these elements. This author still states that the subject analysis of hundreds of documents has shown that the concepts which were found in the speech can be categorized in an unlimited number of categories.

Based on categorizing relations, the concepts which constitute the subjects of a document can be organized to express the relations among them aiming at the representation and organization of information. The semantic relations analysed have been determined from the hierarchical and associative relations.

In order to have a consistent choice of categories and facets of the object in question, we proceeded to the analysis and approach given by the principles of the field of ideas for the categorization of the simplified model. In her simplified model, Spiteri (1998) grouped all the canons, postulates and principles of the field of ideas in two principles, with subdivisions: (1) Principle for the choice of the facet, and (2) Principle for the order of reference of facets and focuses. It is important to highlight that the facet concept has been used in this work.

Among the Principles for choosing the facets and subfacets in the field of ideas of the Simplified Model for Facet Analysis, only the following were considered: **Differentiation, Relevance, Verification, Permanence, Homogeneity, Mutual Exclusivity and Fundamental Categories**. These are principles which come from Ranganathan's FAT, from *CRG*'s Fundamentals or from both.

The **Principle of Differentiation,** proposed by Ranganathan, is based on characteristics of division in which certain common differences or qualities distinguish elements of the same class.

The **Principle of Relevance**, proposed both by Ranganathan and *CRG*, helps to ensure that the chosen facets reflect the proposal, the subject and the scope of the theme in question.

The **Principle of Verification,** proposed both by Ranganathan and *CRG*, states that it is important to choose facets which are definitive and which can be checked. In this case, in which the terminology used comes from the thesis author himself, the principle of verification with a high degree of accuracy in relation to the choice of characteristics has been ensured.

The **Principle of Permanence,** proposed both by Ranganathan and *CRG*, suggests that the chosen facets must represent characteristics of division with permanent qualities in relation to the subject to be divided. The example given previously can also be used as example of this principle, for it reflects the permanence interpretation: the indexer will always be the librarian, which is a permanent characteristic.

The **Principle of Homogeneity,** proposed only by *CRG*, states that the facet must be homogeneous. The contents of two facets must not overlap each other, and each facet must represent only one characteristic of division. For example, when we divide the concept "Text" into the facets "Typology" and "Structure", the terms which appear in one facet cannot appear in the other.

The **Principle of Mutual Exclusivity**, proposed both by Ranganathan and *CRG*, asserts that the facets must be mutually exclusive. Similar to the principle of homogeneity, but with a limiting character, this principle aims to ensure the exclusivity of the classes of an array, deriving from its immediate universe one, and only one, characteristic.

The **Principle of Fundamental Categories,** proposed only by *CRG*, states that there are no categories which are fundamental for the subject as a whole, and that all the categories must be derivative based on the subject to be classified. Thus, this principle can seem opposite to the basic set of Ranganathan's fundamental categories represented by the acronym PMEST (Personality, Material, Energy, Space and Time). *CRG* disagrees with the fundamental categories inasmuch as it prefers to identify them based on the reference of the context of the subject itself, suggesting that no list can be imposed mechanically on the subject. Besides, *CRG* believes that no list should be necessarily exhausting or applied to all the subjects. Louise Spitieri chose the *CRG* approach to be inserted into the *Simplified Model*, bearing in mind that this approach enables the classifiers to better shape the categories for specific subjects, making possible the formation of more distinguished and definite categories. The author also stresses that most systems of faceted/thesaurus classification consulted by her use the *CRG* approach to choose the fundamental categories.

The principles that order the mention of facets and focuses of the field of ideas determine how the focuses will be organized within their respective facets, i.e., how the order of these focuses will be in the array. Among these principles, the **Relevant Succession** and the **Consistent Succession** were used.

In the **Relevant Succession,** Ranganathan as well as *CRG* agree that the order of mention of facets must take into consideration the nature, the subject and the objective of the classification system. This order, which is presented here, derives from the Principle of Relevant Sequence (*Helpful Sequence*) by Ranganathan, and the Principle of Array Order (*Array*) by *CRG*. After the facets were established, we observed the following orders, listed to construct the order with consistency:

Chronological order: increasing chronological order, including operations, which, necessarily, will be carried out one after the other.

Geometric/Spatial order: order for the focuses which have this organizational nature, according to spatial and geometric arrangements.

Order from the simple to the complex: order for elements of growing complexity.

Order from the complex to the simple: order for elements of decreasing complexity.

Canonical order: the most current traditional order, when there is no other principle to follow.

Increase of quantity order: increasing order of quantity, when the subject is in an array of subjects, or the isolated one in an array of isolated ones, has this quality.

Decrease of quantity order: decreasing order of quantity, when the subject is in an array of subjects, or the isolated one in an array of isolated ones, has this quality.

Alphabetical order: alphabetical order by name of a more current international use, when no other sequence of subjects in an array of subjects, or no other sequence of isolated ones in an array of isolated ones is more useful.

The **Consistent Succession**, based on Ranganathan's consistency canon (*Consistent Sequence*), aims to maintain consistency in the structure of a classification system. Its expansion enables the order of mention, of both the facets and the focuses, to be kept consistent.

In order to determine what categories would be used in the thesis, we separated the subjects according to their characteristics. Generally microdocuments do not comprise all the terminology of the subject, and have a limited manifestation of facets. In the case of this thesis, whose theme is "Analysis of subject", which is situated within the subject "knowledge organization", the synthesis of its terminology engendered concepts grouped according to only four categories: Personality, Attribute, Energy and Discipline.

The first determined category was **Personality**, which can also be defined as Entity, and which is considered by scholars, and even by its creator Ranganathan, as one of the most difficult ones to define. The analysis of the entities within the knowledge organization served as reference for this choice, which resulted in the sequence "Librarian", "Authors", "Concept", "Mind", "Document" and "Text".

The second category, **Attribute**, which can also be defined as Matter or Property, was determined based on the concept analysis which characterized or qualified the concepts of the category Personality, described above. Generally the attributes are represented by concepts which refer to abstract properties such as intention, emotion, and abstraction. These concepts represent intrinsic properties: "cohesion", "subjectivity", etc.

The third category, **Energy**, is manifested through actions and operations. As an example of this category in the thesis, we have "general mental processes" (analysis, synthesis, abstraction, etc.), "indexing processes", "subject analysis", "training" and "interpretation". Raghavan (1985, p.29) notices that we must be careful with the following statement: only the situations revealed by intransitive verbs can be manifestations of the category Energy.

Although the thesis studied is focused on the area "knowledge organization", the author Madalena Naves included, in chapter 5, a history of the area of "sciences" and their disciplines and subdisciplines, which justified the determination of the fourth category, **Discipline**.

Then, we proceeded with the distribution of terms according to the principle of the fundamental categories suggested by CRG. Spiteri (1998, p.19) states that "*for fundamental categories all the categories must be derived based on the nature of the subject to be classified*". However, FAT proposes a scheme of categories which, in this case, is considered valid for all the subjects inserted in any area of knowledge.

5. Conclusions

The power of FAT to constitute the logical structure of the hypertext served as the starting point for the modeling of the cognitive structure of the thesis and the analysis of its subject area. The structured approach of FAT made possible the identification of individual characteristics of several concepts, which could be grouped in an analytical way. This enabled the user to see a subject from different viewpoints. This characteristic satisfied the "non linearity" of the hypertext, facilitating the creation of a fluid and interactive structure as the one of the Hypertextual Map – **MHTX**. We can notice that not all the methods for the formation of subjects proposed by Ranganathan were used. Only two out of the five principles proposed by him (dissection, lamination, denudation, reunion/aggregation and superposition) have been used: the methods of dissection and denudation. This occurred due to the fact that the sample, just one thesis, has a specific domain, whose identified concepts were in a specific language, which made it impossible to apply the other methods.

We concluded that the facet analysis technique was efficient in the conceptual modeling of the thesis, providing a dynamic method, from the identification of relevant terms to the formation of categories. The Conceptual Map (CM) provided an alternative approach for the problem of the user's bewilderment, although it did not include the technique of navigation tracking. On the other hand, a navigation guide with a code of colors was created to represent each hierarchical level of the map, showing the user how all the semantic content is organized, and how we connect internally, facilitating a hyperbolical navigation.

References

Barbosa, A. P.(1972).Classificações facetadas.*Ci.Inf*. Rio de Janeiro,Vol.1, No. 2, 73-81.

Bush, V., (1945). As we may think. *Atlantic Monthly*, v.176, n.1, p.101-108.

Campos, M. L. A. (2001). *A Organização de unidades do conhecimento em hiperdocumentos*: o modelo conceitual como um espaço comunicacional para realização da autoria. Rio de janeiro: CNPq/IBICT-URFJ/ECO, 190p. (Tese, Doutorado em Ciência da Informação)

CLASSIFICATION Research Group, (1985). The need for a faceted classification as the basis of all methods of information retrieval. In: CHAN, L. M. et al. (Ed.). *Theory of subject analysis*. Littleton, Col.: Libraries Unlimited,154-167

Duncan,E.B.(1989). A faceted approach to hypertext? In: McALEESE, Ray. *Hypertext:* theory into practice. Nowood, NJ: ABLEX,157-163.

Duncan,E.B.(1989). Structuring knowledge bases for designers of learning materials. *Hypermidia*, Vol.1, No.1,20-33.

Ellis, D.(1996). *Progress and problems in information retrieval*. 2 ed. London: Library Association Publishing, 220p.

Ellis, D. and Vasconcelos, A. (1999).Ranganathan and the Net: using facet analysis to search and organise the World Wide Web. *Aslib Proceedings*, Vol.51, No. 1, 3-10.

Ellis, D. and Vasconcelos, A. (2000).The relevance of facet analysis for World Web subject organization and searching. In:THOMAS, Alan R., SHEARER, James R. *Internet searching and indexing*: the subject approach. New York: The Haworth Press, 97-114.

Glassel, A. (1998) *Was Ranganathan a Yahoo*! Disponible at: *http://scout.cs.wisc.edu/addserv/toolkit/enduser/archive/1998/euc-9803.html*.

Ingwersen,P. and Wormell,I. (1992). Ranganathan in the perspective of advance information retrieval. *Libri*, Vol.42, No.3, 184-201.

Koehler, W. (2001) Concepts. Disponible at *http.//www.ou.edu/cas/slis/courses/LIS5990A/slis5990/catalog/coordination/concepts. htm*

Kwasnik, B. (1992)The role of classification structures in reflecting and building theory. In: AMERICAN SOCIETY FOR INFORMATION SCIENCE, Oct. 1992. Pittsburgh, PA. Advances in Classification Reseach: *proceedings...* Medford: Learned Information, Vol. 3, 63-81.

Piedade, M. R.(1983). *Introdução à teoria da classificação*. 2.ed. Rio de Janeiro: Interciência.

Pollitt, A. S. (1997) Interactive information retrieval based on faceted classification using views. Disponível em http./www.hud.ac.uk/school/cedar/dorking.htm.

Priss, U and Jacob, E. (1999) Utilizing faceted structures for information systems design. In: AMERICAN SOCIETY FOR INFORMATION SCIENCE. ASIS Annual meeting, 62, Oct.31-Nov.4,. Washington, D.C. *Proceedings...* Medford: Learned Information, 1999, 203-212.

Ranganathan, S. R. (1967)*Prolegomena to library classification*. 3ed. London: Asia Publishing House.

Ranganathan, S. R. (1985). Faceted analysis. In: CHAN, L.M. et al. (Eds.) *Theory of subject analysis*: a sourcebook. Littleton, CO: Libraries Unlimited,86-93.

Ranganathan, S. R. (1985).The General theory of classification as the basis for structuring of subject headings. In: REGIONAL CONFERENCE OF THE INTERNATIONAL FEDERATION FOR DOCUMENTATION COMMITTEE ON CLASSIFICATION RESEARCH (FID-CR), 2, 15 Nov. 1985. New Delhi. *Proceedings...* New Delhi: Delhi Library Associations,24-48.

Santos. P. X.(1996). *Engenharia da informação para sistemas de hipertexto*. Rio de Janeiro: CNPq/IBICT-URFJ/ECO, 83p. (Dissertação, Mestrado em Ciência da Informação)

Speziali, P. (1973). Classifications of the sciences. In: DICTIONARY of the history of ideas. New York: Scribners, 462-467.

Spiteri, L.(1998). A Simplified Model for facet analysis: Ranganathan 101. *Canadian Journal of Information and Library Science*, Vol.23,1-30.

Vickery, B.C. (1980). *Classificação e indexação nas ciências*. Trad. Maria Christina Girão Pirolla. Rio de janeiro: BNG/Brasilart