Weblogs Content Classification Tools: performance evaluation

Jesús Tramullas[a,], and Piedad Garrido[a]
[a] *Universidad de Zaragoza, Dept. of Library and Information Science. Pedro Cerbuna 12, 50009 Zaragoza .tramullas@unizar.es*
[b] *Universidad de Zaragoza, Dept. of of Computer Science and Systems Engineering. Ciudad Escolar s/n, 44003 Teruel*

Nowadays, weblogs or blogs are important tools for personal or workgroup websites publication. These tools give the necessary performances to create, edit, evaluate, publish and file digital contents, in the framework of a standarized workflow, and for managing the digital information life cycle. Nevertheless, these tools must be complemented with existence of technical funcionalities necessary to get a correct implantation and use. The aim of the work is to assess the way in which weblogs implement the technical solutions necessary to utilize correctly classification tools. The evaluation took into account let to extract a collection of conclusions of great interest to analize the state of art of the content classification tools integration and the weblogs management systems. As a general conclusion, it can be assured that the current generation of weblogs management systems do not offer all the desired performances for the classical classification tools, offering also a very heterogeneous scene.

Keywords: Classification schemes, weblogs, information retrieval.

1 INTRODUCTION

Weblogs, or blogs, have become a team working tool in many organisations, both formal and informal. Blogs are used to express the writer's opinions or points of view, in the same way as they are being used to promote discussion among other members of their communities. Many organisations have set up a corporate blog which is used as an official meeting point for members, in addition to expounding its objectives, actions or projects. In the field of information units and services, libraries (especially in the Anglo-Saxon world) have been the ones that have most rapidly integrated and developed the capacities of this type of tool, as demonstrated by Bhatt [1].As far as information management is concerned, a blog is a means, it is not a problem, in spite of the problems of proliferation, quality and handling of information, as pointed out by some writers. The problems arise from information illiteracy, which has not been addressed by the education systems over the years, and which blog users will demonstrate as a consequence.

Weblogs, or blogs, are tools containing the features required to create, edit, evaluate, publish and archive digital content within the framework of a standardised flow of work, and to manage the life cycle of this digital information [2]. The main task carried out by blog users is publishing content. A review of the technical architecture and the functionalities of the applications reveals a simple scheme, organised around the publication of content, provided by a simple flow of work. A writer creates the content, assigns some terms to it, (in blog terminology, these are called "categories") describing the content, and publishes it. From then on, the published content/comments come within the reach of all users, who can take part by adding their own comments. A search engine for text information, usually limited to Boolean operators, is incorporated into blog software, as are simple systems for organising the content according to the date of publication (normally a monthly record), or according to categories (by creating a theoretically thematic browsing device). As a key tool for information alerts and dissemination, blog systems have incorporated RSS (which has different meanings, depending on the version, for example, *Really Simple Syndication,* 2.0; *Rich Site Summary,* 0.91 and 1.0). This is a simple content syndication device which can be used by RSS clients, or by content integrators on management systems for more complex information. This article is not concerned with the characteristics and features of RSS.

Research on blogs has centred, in particular, on the study of these from two main perspectives. The first, as an expression of social networks in a digital environment, and the second, from an analysis of the links they provide as a means of retrieving and accessing information, especially through the use of methods and techniques of trawling the web. Linked to both these approaches is the growing interest that independent classifications, called *folksonomies,* created by the users, have generated among the specialists, given the expansion reached by services such as Technorati, Flick.er, and Youtube.

2 USERS AND CLASSIFICATIONS

The availability of simple tools for publishing content has made it possible for users to publish

information quickly in digital environments, without the need for advanced technical know-how. However, most users of these environments lack experience in description, tagging and metadata for digital information. Consequently, and given that the exchange and access to the information depends on the descriptive tags used, called "categories" in tools for blogs, users have started to use descriptive terms, or "categories" totally based on intuition.. In this environment, and because of the heterogeneity of the categories, services such as Technorati use classifications based on the principle of the popularity of the categories, or *folksonomies*, which are represented visually by "tag clouds".

The use of classifications for content has been widely studied by information and documentation science. Mai [3] has argued that bibliographic classification can contribute a large step forward in the organisation of web content, providing that research into this fits in with the characteristics of the digital environment, and is centred on classifications based on information needs and uses. However, classifications are only being used systematically in specialised contexts. Methods for creating faceted environments have aroused most interest in corporate environments (intranets) and e-commerce. Specific applications for these can be seen in product catalogues, browsing menus, etc. As an example, you need look no further than the references available in IAwiki [4] and in IAslash [5], or Denton's [6] bibliography review.

Garshol [7], in his review of classification schemes and metadata for application on the semantic web, highlights the difference between controlled vocabulary, taxonomy (or hierarchical classification), thesaurus, faceted classification, ontology and topic map. From the experience of digital information users, it can be said that the first two are the most widely used, because of their simplicity, for such things as virtual communities, forums and blogs. However, the more specialised information services have opted for faceted classifications and thesauri, while ontology and topic maps continue to be used mainly in very advanced, high cost environments and products, and in academic research.

The use of the principles of faceted classification has been analysed by Zins [8], who studied various models of organising information in web portals. He identified a total of eight faceted classification models, relating to subjects, objects, applications, users, localisation, references, media and languages. There are models that combine several of these. The first five refer to the information content, which is the most widely used approach in digital information services and portals.

3 EVALUATING FUNCTIONALITIES FOR CLASSIFICATION

The use of different types of classification schemes in blogs depends, on one hand, on the knowledge and skills of the users and, on the other hand, on the technical functionalities implemented in the blogs. The aim of this article is to evaluate the way in which weblogs implement the technical solutions required to use classification tools correctly. To this end, the following evaluation process has been used:

1. Defining the required functionalities
2. Selecting the weblog management systems to be evaluated
3. Empirical testing of the available functionalities
4. Synthesising and validating the results

The definition of functionalities derives from the classification schemes that are under evaluation. For this study, items under evaluation are considered to be simple key words, hierarchical classifications or taxonomies, faceted classifications, thesauri and ontology, in accordance with Garshol [7]. Therefore, an analysis was made to see if the software packages for blogs incorporated, firstly, support for generic categories and, secondly, if the available functionalities enabled more advanced classification schemes to be implemented.

This type of tool with its limitations has been around for a long time in the field of tools and functionalities for classifying content. For example, in 2003 Rabourn [9] had already developed a solution able to use faceted classifications in *Movable Type*, through the use of regular expressions (*regexp*). There is a very large panoply of software for blogs, and this can include both the packages available to the end user to obtain and install, and the free or paying services available through a provider, such as Blogia, Blogger, TypePad, etc. Two criteria were established in order to choose the software for evaluation. In the first place, they had to be full packages distributed with recognised freeware licences. In the second place, their standard installation had to include functionalities for content classification. A check of the software available led to six being selected: WordPress, Serendipity, TextPattern, LifeType, b2evolution and NucleusCMS. All the tools are Type II, as proposed by Du and Wagner [10], and are well established on the Internet, with development and user groups, thanks to whom there are a great many weblogs that have implemented these

solutions.

Once these were all installed with the basic set-up, several inputs of content were created in the various blogs, with the categories being assigned to them corresponding to the different classification schemes mentioned in above sections. In fact, nearly all the tools required the categories to have been entered before the input, as these are selected from the screen for creating and editing content. The results obtained are shown in table 1.

Table 1. Analysis of software packages for blogs

Blog	***Vers.***	***URL***	***Observations***	***Additional plug-ins***
WordPress	2.0.3	www.wordpress.org	Enables direct selection by category from the side menu. Categories from different hierarchical levels can be combined.	WordPress is the package offering the greatest variety of plug-ins to improve the system of content description. However, most of these are for keywords, and do not seem to support more advanced classification schemes.
Serendipity	1.0	www.sys9.org	Enables direct selection by category from the side menu. Categories from different hierarchical levels can be combined.	*Category Tree Menu*
TextPattern	4.0.3	www.textpattern.com	Enables the content to be organised by section (single) and by category (a limit of 2) *Keywords* can be included in the advanced editing.	*rss_unlimited_categories* *dak_categories* *asy_category*
LifeType	1.0.5	www.lifetype.net	Only simple categories can be used.	
b2evolution	0.9.2	www.b2evolution.net	Enables direct selection by category from the side menu. Includes a more precise text search. Plug-ins can be installed to improve the use of categories in browsing and text search.	*Expandable categories* *SEO Keywords & Description* *Category Behavior*
NucleusCMS	3.23	www.nucleuscms.org	Only simple categories can be used. The standard installation does not allow the use of multiple categories, additional plug-ins are required.	*NP_Keywords* *NP_Multiplecategories*

In fact, the functionalities for implementing classification schemes, from the point of view of an expert in information, are limited. The basic use carried out by users is merely assigning flat categories through keywords. Table 1 contains a column showing the additional modules (*plug-ins*) which can be installed to increase the features of classification schemes. For example, with WordPress, most of the available plug-ins have been designed to use keywords, not classification schemes. It should be pointed out that some facets, such as author or date of publication, are not included in the classification functionalities. However, the packages provide the opportunity to open the content using both, by enabling the plug-ins in the set-up

options. In this way, blogs can provide browsing devices and points of access in their basic set-up for three facets: information content, author and date, although these cannot be combined, which is a serious limitation.

This limitation of possible combinations for selecting input also extends to the internal text search engines. All the packages analysed use text search features provided by the MySQL relational database management system. They do not provide interrogation interfaces to enable precise searches beyond the presence of the terms, and only b2evolution offers the basic Boolean and sentence search.

Table 2. Functionalities for classification schemes.

Blog	*Keywords*	*Hierarchical Class. /Taxonomy*	*Faceted Class.*	*Thesaurus*	*Ontology*
WordPress	**X**	**X**	**X (incomplete)**	**-**	**-**
Serendipity	**X**	**X**	**X (incomplete)**	**-**	**-**
TextPattern	**X**	**-**	**-**	**-**	**-**
LifeType	**X**	**-**	**-**	**-**	**-**
b2evolution	**X**	**X**	**X (incomplete)**	**-**	**-**
NucleusCMS	**X**	**-**	**-**	**-**	**-**

Table 2 shows the types of classification schemes supported by packages for blogs. Three of these, WordPress, Serendipity and b2evolution, are clearly superior and also offer plug-ins to boost their features. However, their functionalities remain limited and this also shows up in the devices provided for exploring and retrieving information. There is an interesting way to develop better classification schemes in blogs using the basic features of categories. This is what Hearst [11] called hierarchical faceted categories, made up of a set of category hierarchies, each of which corresponds to a different facet. This approach has been tested through the *Flamenco* prototype.

Fig. 1. b2evolution: Categories management

4 CONCLUSIONS

Du and Wagner [10] have analysed the relationship between a successful blog and its technical characteristics. The results obtained show that success depends on the value offered to users, on the technological features and the social levels of the relationship. However, with the technological aspect, they place the emphasis on the functionalities for the relationship with other blogs and the creation of virtual communities. This approach is predominant in the context of working with blogs. In the first place, it must be pointed out that previous evaluation criteria place greater importance on publishing content, on presenting the end user, and the life cycle, than on classification, exploration or information retrieval functionalities. Consequently, in second place, the functionalities available for implementing classification schemes are highly heterogeneous.

In fact, the end users do not seem to mind this limitation, and content themselves with using keywords. However, this technical limitation contrasts with the development and evolution of social systems of tagging the content, or folksonomies, which are becoming the standard system for classifying content, as proved by the success of tag clouds. Blogs are simplifying the use of classification schemes on the web, giving precedence to schemes based on the popularity of simple descriptors, without hierarchies or relationships. Not even the Dublin Core, with all its simplicity, has been accepted for the content description of blogs, apart from certain points. If Web 2.0 has one of its supports in the ontologies, it could turn out to be surprising that the best developed tool for publishing content in the last few years is being almost completely sidelined, or else it could be understood that there is, in fact, a long way to go before reaching Web 2.0.

Neither has the semantic web been integrated into blogs yet, although this can be done through the development of new functionalities. Karger and Quan [12] have developed a prototype of a client for a semantic blog, based on Haystack, which uses RFD to tag the information content, while at the same time using ontologies, through OWL, to add semantics to RSS. The semantics are included through a class specially designed to reflect the categories. The prototype uses XSLT to obtain semantic tagging. This approach is also described by Cayzer [13], using a semantic tool for blogs which automatically generates metadata tagged in RDF.

REFERENCES

[1] Bhatt, J. 2005. Blogging as a Tool: Innovative Approaches to Information Access. Library Hi Tech News incorporating Online and CD Notes, 22(9), pp: 28-32

[2] McBride, M. 2004. Open source weblog and content management systems for the information professional. *Searcher* 12 (9), pp: 24-29

[3] Mai, J-E. 2004. Classification on the Web: Challenges and Inquiries. Knowledge Organization, 31(2), pp: 92-97.

[4] IAwiki: Faceted Classification. 2006., http://www.iawiki.net/FacetedClassification

[5] IAslash: Faceted Classification 2004. http://www.iaslash.org/taxonomy/term/101

[6] Denton, W. 2003. Putting Facets on the Web: An Annotated Bibliography. http://www.miskatonic.org/library/facet-biblio.html

[7] Garshol, L.M. 2004. Metadata? Thesauri? Taxonomies? Topic Maps! Journal of Information Science, 30(4), pp: 378-391

[8] Zins, C. 2002. Models for Classifying Internet Resources. Knowledge Organization, 29(1), pp: 20-28.

[9] Rabourn, T. 2003 Faceted MovableType http://www.pixelcharmer.com/fieldnotes/archives/process_designing/2003/000348.html

[10] Du, H.S., Wagner, C. under publication. Weblog success: Exploring the role of technology. International Journal of Human-Computer Studies

[11] Hearst, M. A. 2006. Clustering versus faceted categories for information exploration. Communications of the ACM, 49(4), pp: 59-61

[12] Karger, D.R., Quan, D. 2005. What would it mean to blog on the semantic web? Journal of Web Semantics, Web Semantics: Science, Services and Agents on the World Wide Web, 3. pp:147–157

[13] Cayzer, S. 2004 Semantic Blogging and Decentralized Knowledge Management. Communications of the ACM, 47(12), pp: 47-52