Classification and Indexing of Web Content Based on a Model of Semantic Social Bookmarking

Antonello Angius[1], Giulio Concas[2], Dino Manca[3], Filippo Eros Pani[2] and Georgia Sanna[3]
[1]*Centro Regionale di Programmazione, Regione Autonoma della Sardegna, via Battisti, Cagliari, Italy*
[2]*Department of Electrics and Electronics Engineering, University of Cagliari, Piazza d'Armi, Cagliari, Italy*
[3]*Experteam srl, Via Zara 11, Cagliari, Italy*

Keywords: Multimedia Content, Social Networks, Social Bookmarking, Taxonomy, Folksonomy.

Abstract: One of the key challenges in Information Technology is finding a way to organize the knowledge present on the Web. This led to years of research on the integration of information, on the Semantic Web and related technologies. Information Search and Retrieval from the Web occur through a process of content disambiguation and search engines use algorithms and software agents in order to meet the needs of users and advertising buyers. Ex-post analytical agents and tools are becoming more pervasive, so much to cause increasing problems of privacy. Our work proposes an innovative approach of content disambiguation that overturns the ex-post semantic analysis of contents, because it deals with an ex-ante classification conducted on two axes: vertical one (hierarchical and taxonomic axis) and horizontal one (folksonomic axis through tags or keywords). This method, which is based on the logic of social bookmarking and focuses on semantic tagging, represents a new frontier in information architecture because it introduces a new way of classification made by people using keywords that have a specific lexical and semantic value. This approach will allow people to create a knowledge base of Web contents characterized by a precise semantic definition.

1 INTRODUCTION

Search engines use several software agents and original algorithms to meet needs of private users and advertising buyers. However, in order to achieve satisfying results in searches through “words”, information on the Web should be semantically connoted.

On the user's side, it is often difficult to filter properly the information through thousands of search results, while on the advertising's side there is a relevant percentage of "contextual" messages that are outside the target and therefore they are ignored or unwelcome by people.

The approach proposed in this work is based on a technique of disambiguation that overturns the ex-post semantic analysis of contents, because it deals with an ex-ante classification conducted on two axes: vertical one (hierarchical and taxonomic axis) and horizontal one (folksonomic axis through tags or keywords).

The basic idea of this approach is to allow people to classify contents with an innovative approach related to social bookmarking: users will be enabled to bookmark a content, classify it semantically and add it to the knowledge base.

According to the logic of social bookmarking, the knowledge base will be enriched/supplemented by other users with further classified contents. This method will promote the development of a new collaborative model of Web 2.0, where people create the knowledge base and classify interesting contents drawn from social networks, other UGC (User-Generated Content) or Web sites (Lunesu et al., 2011) (Lunesu et al., 2011).

The paper is structured as follows: Section Two shows an overview of the knowledge bases used to classify contents. In Section Three we present the most relevant components of the system (plugin, classification model, methods of interaction among users). Section Four includes the conclusions and some reasonings about our work. Section Five deals with some possible future developments.

Angius A., Concas G., Manca D., Eros Pani F. and Sanna G..
Classification and Indexing of Web Content Based on a Model of Semantic Social Bookmarking.
DOI: 10.5220/0005139303130318
In *Proceedings of the International Conference on Knowledge Management and Information Sharing* (KMIS-2014), pages 313-318
ISBN: 978-989-758-050-5

2 KNOWLEDGE BASES USED FOR CLASSIFICATION

Starting point of this work is the examination of baseline studies and tools, including UGC, which are used to disambiguate and classify word meanings, in order to categorize contents with disambiguated terms (Passant and Laublet, 2008) (Suchanek et al., 2008) (Zhong, 2008) (Agirre and Soroa, 2009) (Murgia et al., 2010).

Consequently, in this section we describe the major scientific lexical-semantic projects such as WordNet and BabelNet, which assurance a correct disambiguation of terms at the lexical and semantic level, and resources which are collaboratively run by largely volunteers according to the logic of Web 2.0 (Berners-Lee et al., 2001), such as the well-known Wikipedia and DBpedia project.

Among the other interesting projects on tags meaning disambiguation and RDF metadata processing, we mention PiggyBank, which is a tool integrated into the Web browser that allows users to extract information items from within Web pages and save them in Semantic Web format, replete with metadata (Huynh et al., 2005) and DogmaBank, which is a social bookmarking tool that can be used to tag Web pages with freely chosen keywords or with concepts. (Spyns et al., 2006).

2.1 WordNet

The proposed technique is based on WordNet, a large semantic-lexical database of English, whose development began in 1985 at the Cognitive Science Laboratory of Princeton University under the direction of linguists and psychologists.

WordNet aimed at developing a more intuitive dictionary/thesaurus and a support for Artificial Intelligence applications and for technologies related to automatic text analysis. It groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets) by means of semantic relations (between word meanings) and lexical relations (between word forms). Its database includes 117000 synsets, which are interlinked by means of a small number of conceptual relations, contains a brief definition and, in most cases, one or more short sentences explaining the use of synset members. (Miller, 1995).

WordNet differs from a common thesaurus firstly because it creates relationships not only between word forms (strings of letters) but also between the specific sense of words. Secondly, it labels the types of lexical or semantic relationship among words, whereas a common thesaurus groups words just on the level of meaning similarity (Fellbaum, 1998).

2.2 Wikipedia

Wikipedia, the well-known universal "free encyclopedia that anyone can edit", well represents the spirits in which free software is developed: its keywords could be collaboration, sharing and improvement.

It is run largely by volunteers and it is not regulated by a central committee. However, there are review processes: although everybody can contribute, regardless of the education or expertise level, the user is warned about the possibility of change/deletion/redistribution of contents, where the last possibility depends on the Creative Commons Licence.

Wikipedia articles, represented as pages (WikiPages), contain information on a concept (such as *Play-Theatre*) or on an entity (like *William Shakespeare*). Each WikiPage's title includes the concept's term (*Play*) in addition to an optional label enclosed in parentheses that, in cases of ambiguity, specify its meaning (Wu and Weld, 2007) (Adar et al., 2009) (Toral et al., 2009).

2.3 DBpedia

DBpedia project focuses on converting Wikipedia information into structured knowledge, which is then made available on the World Wide Web as RDF (Resource Description Framework), and allows to execute complex queries on Wikipedia contents and to link the extracted data to other datasets (Lehmann et al., 2014) (Morsey et al., 2012). It presents the following features:

1) a framework for data mining, which converts Wikipedia's content into RDF (Brickley et al., 1999);
2) Wikipedia's content is available through a huge and multi-domain RDF dataset;
3) the dataset can be connected to other datasets with a resulting Web of data that contains, in all, 2 billion RDF triples (Garcia et al., 2009).

2.4 BabelNet

BabelNet is a semantic network in constant evolution, which includes encyclopedic and lexicographic items. It has been developed by Roberto Navigli, Paolo Ponzetto (and others) at the University of Rome "La Sapienza" within the "ERC Starting Grant MultiJEDI MultiJEDI" (Navigli and Ponzetto, 2010).

Main aim of BabelNet is to provide a unified resource which integrates Wikipedia and WordNet through a system of automatic mapping: WordNet furnishes the missing relationships in Wikipedia and Wikipedia adds the multilingual dimension.

Multilingualism is another purpose of BabelNet: the lexicalized concepts belonging to the different languages are collected from Wikipedia through interlingual links and, in case of poor languages, lexical gaps are filled with the aid of automatic translation systems.

Similarly to what has been proposed by WordNet, in Babelnet words are grouped into sets of synonyms called Babel synsets. Each Babel synset provides a given meaning and contains all the synonyms which express that meaning in several languages (Navigli, Ponzetto, 2012).

BabelNet 2.5 (May 2014) is obtained from the automatic integration of six linguistic resources: WordNet 3.0, Wikipedia, Open Multilingual WordNet, OmegaWiki, Wiktionary, Wikidata.

3 THE PROPOSED APPROACH

The proposed approach is based on an ex-ante classification conducted on two axes: vertical one (hierarchical and taxonomic axis) and horizontal one (folksonomic axis through tags or keywords). This technique aims at being used by people who want to classify semantically information items and add them to the bookmark. All the contents classified by individual users will be categorized in a large knowledge base, that could be shared with other users.

According to the logic of social bookmarking and to a new collaborative model of Web 2.0, other users will enrich and integrate this knowledge base by classifying other contents, which could also be shared and commented on social networks, forums and blogs (Concas et al. 2008). People will be able to use this tool to find contents, searching them by means of their classification or their disambiguated tags.

The set of pre-classification services and results matching is run by a personal classification engine. This approach involves a main axis of innovation which consists of a new model of information classification and which includes two integrated types of skills related to two different fields: on one side the information architecture and on the other side the software.

On the side of information architecture, we are developing a new structure of thesaurus (a linguistic database that contains logical, semantic and hierarchical relations among terms) based on BabelNet.

This tool aims to be easily used and integrated by people to classify their contents of interest on a horizontal axis (in which tags refers to the folksonomy based on the thesauri) and on a vertical axis (regarding the hierarchical-taxonomic assigning of tags), with a resulting semantic disambiguation of the classified contents.

3.1 Plugin Software

From a software point of view, the system allows users to select through a specific plugin a content on a social network (such as Facebook) or on any Web site. The main CMS environments (e.g. Drupal, Joomla!, Wordpress,) for Social Networks (Facebook, Twitter, etc.) and forums are under development.

In addition, we have developed a bookmarking toolbar that enables users to select areas of text and images which can be classified and added to the bookmark, with the purpose of making them available on the knowledge base.

New features of social semantic bookmarking will be provided by this plugin, with the added value of a total ex-ante disambiguation, whose functional and social implications are quite extensive. It is therefore evident that our work does not provide a "simple" way of bookmarking contents which integrates or it is alternative to classic old social bookmarking systems such as Del.icio.us, which is a social bookmarking Web service for saving, organizing and discovering links on the Web, that can be used to tag information items with freely chosen terms. disambiguation of the classified contents.

3.2 Relationships among Users

The system enables to develop a knowledge base which will be different from those provided by common search engines because it will be based on an ex-ante classification, socially controlled, and therefore more precise than those provided by ex-post classification. It will also allow new ways of relationships among users, supporting the communication and the exchange of information through new forms of interaction based on common thematic interests.

This new technology will be completely different from the traditional "friendship on Web" model, called FOAF (Friend Of A Friend), which is

normally developed through concatenation processes. People will receive notifications whenever a content pertaining to their area of interest is published, supporting therefore the interaction with those people who have inserted the content into the knowledge base.

This new form of interaction, depending on thematic basis, is also different from that of "groups", which is typical of social networks and some forums, because in our system the first contact among users does not depend on a general adherence to themes, titles or designations of groups.

The technique of disambiguation is activated by the personal interests classified by the user through a built-in feature, in order to automatically perform searches and daily crossings with all the informative items having the same or similar classification.

This means that once people set a "dashboard" of contents and personal interests (freely configurable and upgradeable), they will be able to view at each access all the contents and information inserted by other users which are consistent to their dashboard. Consequently, people will no longer have to search information and contacts.

3.3 The Classification Model

On the side of information architecture the classification model, which aims at the pre-indexing of content made by users, is developed on the basis of the latest scientific and empirical studies. Our work focuses on the development of a rapid method of classification, which can be easily used by people though an appropriate interface and its tools. The system enables the user to choose, for each tag which is associated to a content, one of the semantic meanings drawn from BabelNet. Inserting the tag “apple”, for instance, the user can select from the tags menu the most appropriate meaning for his contents, choosing one of the following senses: *«The apple is the pomaceous fruit of the apple tree, ... etc.», «The Apple Retail Store is a chain of retail stores owned and operated by Apple Inc., dealing in computers and consumer electronics», «The apple is ..»*.

It is clear how this type of classification, centered on the *word meaning* in place of the *word form*, is particularly useful in the cases in which a content is classified with polysemantic tags, connoted with more than one meaning.

In addition to the assignment of semantic tags, each content is classified by people with one category and its subcategories.

Content categorization can be done through two modalities: user can choose one of the *Suggested Categories* proposed by the Software, which are semantically related to the tags inserted, or freely select a tag and its sub-categories from the dedicated menu. Categories are extracted from the Open Directory Project, also known as Dmoz, which classifies Web sites into 16 main categories, each of which is divided into several sub-categories.

In addition to the above classification model, each content can be classified through a meta-data, also called faceted attribute. The system identifies four main meta-data, which meet the primary information needs of user: *“ask or give information”, “test and reviews”, “How to and guides”, “Exchange, sell or buy”*.

The figure below shows the logical components of the system proposed: the user employs the bookmarklet to select interesting content (texts, images, videos) on the World Wide Web; then he logs in to our Web site in order to classify semantically the selected content with tags and categories extracted from the knowledge bases.

Furthermore, the system will allow people to share the classified content on the main social networks, such as Facebook and Twitter, supporting new forms of social crossing interaction and new ways of relationships among users, based on common thematic interests.

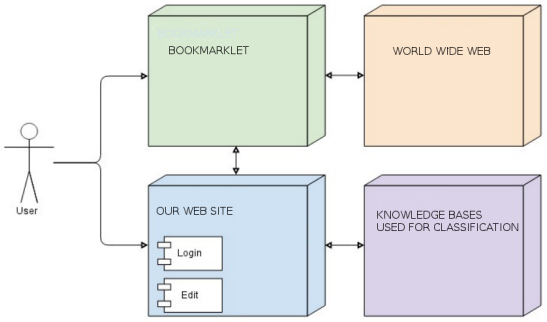

Figure 1: The logical components of the system proposed.

4 CONCLUSIONS

Our proposal introduces a new method of content classification, based on an pre-classification technology that allows an ex-ante disambiguation which is significantly more reliable than the current ex-post indexing techniques.

This will allow users to create both a personal semantic knowledge base, a kind of user's personal

bookmark, and a collaborative knowledge base created by the semantic social bookmarking.

The new service is particularly attractive for the user who feels the need for effective information and wants to classify his interests, using a simple and intuitive tool that suggests a disambiguation taxonomic tree dealing with the typed tags (with the possibility of reverse path, from general to specific) (Pani et al., 2012) (Pani, et al., 2013).

Once the content classification is ended, the service allows user to search and match all the similarly classified information, which can be filtered according to optional criteria.

It is a technique which has a high potential for innovation and market penetration and which will form the basis for the implementation of an innovative system which will enable users to:

- select and file any Web content (texts, videos, images);
- create an archive of content in the personal area;
- meet other community users on the basis of shared interests and passions;
- find content through the search by title, category and / or keywords;
- share contents on social networks (Facebook and Twitter) and interact on them simultaneously through the innovative social crossing.

5 FUTURE DEVELOPMENTS

The possibility of obtaining an ex-ante disambiguation of Web content allows users to communicate and direct advertising with greater preciseness, opening an interesting frontier of innovation in this field too.

Thanks to the folksonomic and taxonomic classification of contents, a more precise contextual advertising will be achieved. It is well known that the best advertising addressing induces a more effective use, and it can become an effective service when the message is characterized by objective information elements, and/or interactive inputs, and/or types of promotional benefits to the user.

Our idea pertains to two main fields:

- to the social and knowledge field, because it proposes a management tool updated to the real interests of everyday life, with characteristics of interoperability between different systems and Web communities;
- to the economic field because, thanks to its high disambiguation accuracy, it enables to better address advertisements and other paid services, which will be potentially appreciated by people. information and contacts.

5.1 Further Social Features

In the light of these considerations, our technical proposal belongs to the social area, which includes social networks, thematic forums and information sites whose tools support forms of interaction and communication among users.

Baseline studies on qualitative and quantitative dynamics, at global and regional levels, of the major social networking services reveal a less explosive phase of growth than in the past, with the consolidation of the most popular players and "winners" in the marketplace.

However, at the same time relevant dynamics of change are occurring in the quantity and types of user activities, as a result of innovations and new features incorporated in the main services and in the market for plugins and third-party applications.

REFERENCES

Moore, R., Lopes, J., 1999. Paper templates. In *TEMPLATE'06, 1st International Conference on Template Production*. SCITEPRESS.

Smith, J., 1998. *The book*, The publishing company. London, 2nd edition.

Adar, E., Skinner, M., Weld, D. S., 2009. Information arbitrage across multi-lingual Wikipedia. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, Barcelona, Spain, pp. 94–103.

Agirre, E., Soroa, A., 2009. Personalizing PageRank for Word Sense Disambiguation: In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, pp. 33-41.

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The Semantic Web. In: *Scientific American*, pp. 29-37.

Brickley, D., Guha, R. V., 1999. *Resource Description Framework (RDF) Schema Specification*. Proposed Recommendation, World Wide Web Consortium. http://www.w3.org/TR/PR-rdf-schema

Concas, G., Lisci, M., Pinna, S., Porruvecchio, G., Uras, S., 2008. Open Source Communities as Social Networks: an analysis of some peculiar characteristics. In: *19th Australian Conference on Software Engineering, ASWEC 2008*, pp. 387-391.

Creative Commons, http://creativecommons.org

DBpedia, http://it.dbpedia.org

Delicious, https://delicious.com

DMOZ (Open Directory Project), http://www.dmoz.org

Drupal, https://drupal.org

ERC Starting Grant MultiJEDI, Università di Roma "La Sapienza". http://lcl.uniroma1.it/multijedi

Facebook, https://www.facebook.com

Fellbaum, C. (Ed.), 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Garcia, A., Szomszor, M., Alani, H., Corcho, O., 2009. Preliminary Results in Tag Disambiguation using DBpedia. In: *Knowledge Capture (K-Cap'09)*, First International Workshop on Collective Knowledge Capturing and Representation, CKCaR '09, Redondo Beach, California, USA.

Huynh, D., Mazzocchi, S., Karger, D., 2005. Piggy bank: experience the semantic web inside your web browser. In: *Proceedings of the 4th international conference on The Semantic Web*, Galway, Ireland.

Joomla!, www.joomla.org

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. In: *Semantic Web*, IOS Press, ISSN: 2210-4968.

Lunesu, M. I., Pani, F. E., Concas, G., 2011. An Approach to manage semantic informations from UGC. In: *3th International Conference on Knowledge Engineering and Ontology Development, KEOD 2011*, Paris, France.

Lunesu, M. I., Pani, F. E., Concas, G., 2011. Using a standards-based approach for a multimedia knowledge-base. In: *3th International Conference on Knowledge Management and Information Sharing, KMIS 2011*, Paris, France.

Miller, G. A., 1995. WordNet: A Lexical Database for English. In: *Communications of the ACM*, Vol. 38, No. 11, pp. 39-41.

Morsey, M., Lehmann, J., Auer, S., Stadler, C., Hellmann, S., 2012. DBpedia and the live extraction of structured data from Wikipedia. In: *Program: electronic library and information systems*, Vol. 46 Issue 2, pp.157-181, ISSN: 0033-0337.

Murgia, A., Concas, G., Marchesi, M., Tonelli, R., 2010. A machine learning approach for text categorization of fixing-issue commits on CVS. In: *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*.

Navigli, R., Ponzetto, S. P., 2010. BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 216-225, Stroudsburg, PA, USA.

Navigli, R., Ponzetto, S. P., 2012. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. In: *Journal Artificial Intelligence*, Elsevier Science Publishers Ltd. Essex, UK, Vol. 193, pp. 217-250.

OmegaWiki, http://www.omegawiki.org

Open Multilingual Wordnet, developed by Francis Bond, Nanyang Technological University, Singapore, http://compling.hss.ntu.edu.sg/omw

Pani, F. E., Lunesu M. I., Concas, G., Stara, C., Tilocca, M. P., 2012. Knowledge Formalization and Management in KMS. In: *4th International Conferenze on Knowledge Management and Information Sharing, KMIS 2012*, Barcelona, Spain.

Pani, F. E., Lunesu, M. I., Concas, G., Baralla, G., 2013. An Approach to Manage the Web Knowlledge. In: *5th International Conference on Knowledge Engineering and Ontology Development, KEOD 2013*, Algarve, Portugal.

Passant, A., Laublet, P., 2008. Meaning Of A Tag: a collaborative approach to bridge the gap between tagging and Linked Data. In: *Proceedings of the Linked Data on the Web Workshop (LDOW2008) at the 17th International Semantic Web Conferences (ISWC 2008)*. Karlsruhe, Germany, ISSN: 1613-0073.

Spyns, P., de Moor, A., Vandenbussche, J., Meersman, R., 2006. From folksologies to ontologies: how the twain meet. In: *Proceedings of the 2006 Confederated international conference on On the Move to Meaningful Internet Systems: CoopIS, DOA, GADA, and ODBASE*, Vol. 1, pp. 738-755, Springer-Verlag Berlin, Heidelberg. ISSN: 3-540-48287-3 978-3-540-48287-1

Suchanek, F. M., Kasneci, G., Weikum, G., 2008. YAGO: a large ontology from Wikipedia and WordNet. In: *Journal of Web Semantics*, Vol. 6, pp. 203–217.

Toral, A., Ferrández, O., Agirre, E., Muñoz, R., 2009. A study on linking Wikipedia categories to WordNet synsets using text similarity. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 449-454 .

Twitter, https://twitter.com

Wikidata, http://www.wikidata.org/

Wikipedia, http://www.wikipedia.org

Wiktionary, http://www.wiktionary.org/

WordNet, http://wordnet.princeton.edu

Wordpress, https://wordpress.com

Wu, F., Weld, D., 2007. Automatically semantifying Wikipedia. In: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Managemen t*, Lisbon, Portugal, pp. 41-50.

Zhong, Z., Ng, H. T., Chan, Y.S., 2008. Word Sense Disambiguation using OntoNotes: An empirical study. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, Hawaii, pp. 1002-1010.