Meta data-based Text mining for Web Content Categorization

S.Sarasvady
Amrita Vishwa Vidyapeetham University
Ettimadai
Coimbatore - 641 105. TN
India
saras_vady@yahoo.com

ABSTRACT**:** *In the recent period web resources gain momentum in terms of extensive refined processing resulting the inclusion of a very large number of pages retrieval for any given query. Building taxonomy of the source becomes essential in order to deal with the process of related divisions and terms. The web world has many systems and architectures for processing content-based or context-based features of Source. In this paper we analyze these developments and attempt to generate a set of features to improve web page taxonomy of selected source by classifying web pages as main and fractals. We did a series of testing using an extensive datasets of Source by deploying both using content-based and context-based web page features. Our testing and implementation ensure a high level success in the combination of web content analysis.*

Keywords: Web page classification, Web taxonomy, Metadata

Received: 18 July 2009, Revised 28 August 2009, Accepted 16 Sepetmber 2009

1. Introduction

Producing metadata is the act of creating standardized description of the content. Web research can be optimized for good quality in an efficient manner if web resources are organized and created in relation at the metadata level.

In the recent years the success of digital libraries, the sustenance of interoperability, the Open Archives Initiative and the evolution of Semantic Web all rely on efficient metadata generation. This paper is the result of the generation of landscape for metadata framework which includes the actions or work, based on a content similarity measurement tool.

In the last decade, metadata leads to a new look with the deployment of similarity measurement as a platform and creation of a standard web content. It is known that the content similarity as a language provides a new versatile structure for identifying and fixing metadata as the rapid proliferation of digital content demands both rapidly produced descriptive data and the encoding of more types of metadata. In the recent years there is an emergence of many content measurement systems with aim to harness these developments for web content optimization. A major initiative is the creation and introduction of the Metadata Object and Description Schema (MODS)1, a MARC-compatible XML schema for encoding descriptive data. Another related development is the Metadata Encoding and Transmission Standard (METS)2, a highly flexible XML schema for the inclusion of the descriptive metadata and various other important types of metadata required to generate the standards web research.

2. Taxonomy of Web Content

The simplest approach for web page classification is only using the text features. However, Source are more than texts, and

they contain a lot of context and structural features, e.g., links, anchor texts, URLs, etc. These features can be divided into two broad classes [3].

a) The first group is called as the On-page features, which are directly located on the page to be classified. These features include the following.

- Textual content and the tags assigned that is the most straightforward feature.

- Term analysis. Term representation rendered by a test document and provides different views of users for similarity measure. It is evident that the content detection in the text relies on the ways of assigning tags, using similarity information of the rendered page. This is clearly more generic than analyzing document structure focusing on HTML tags, because different tagging may have the same rendering effect. We propose the On-page features as *Main*.

b) The second one relies on the links which accommodate the features of relations, which are found on the pages related in some way with the page to be classified. The main On-page features are useful but they provide information only from the viewpoint of the page creator. Sometimes, it is necessary to use features that do not reside on the page. We in this paper propose the links and relations as fractals.

The major issue in web page processing the elimination of many irrelevant, infrequent and stop words that reduce the performance of the classifier, extracting or selecting representative features from the web page which is the requirement in web content processing system.

There are a number of possible approaches in information management that require discussions and applications on taxonomies: [4]

Any piece of semantic content needs metadata on structured data as the basis for quantitative analysis, taxonomy on unstructured content for the results of analysis. The reality in the present web content is concept extraction from content. We can extend traditional quantitative web content extraction with qualitative extraction from unstructured content once a taxonomy is applied. For example:

• analysis of risk/no risk claims by category
• analysis of call center issues by product to monitor a recall
• analysis of content effectiveness by social media context

Now the semantic content looks at the role of taxonomy in Data Management

• The taxonomy should eliminate structured data redundancy
• Taxonomy should eliminate unstructured content ambiguity

The tools can be build up in the sense that it can ascribe meaning and intent as humans think in ambiguous terms. Information systems would need to know something about the perspective, experience and objectives in order to precisely determine the correct information to present. Though tools and technologies are getting better, they are not yet at the point where domain specific expertise and human judgment are not needed.

We look at a broad scope for taxonomy and problems when unstructured content needs to be classified. There are many illustrations to show and document the ways to mine consistent causes and effects from unstructured notes.

Thus building taxonomy has strong scope as we can build hierarchical and associative taxonomy (in contrast to other controlled vocabularies or thesauri), every term is expected to have the hierarchical as well as associative relationships, but not every term must have an associative relationship. Such relationships should only be created as needed to point to related terms that might not otherwise be known to exist to the user.

Thus, we have a hierarchical taxonomy, 2 levels and we understand that every term must have a hierarchical relationship but not everyone needs an associative relationship. [5] We are also not looking to establish relationships between terms within

the same facet. The concept and relation that is driving to this path is the opportunity to expose related content through those relationships.

We provide a conceptual example: In a target web page, we may create a tag-driven Topic Page on Icing. On this page then we would want to expose related content via dynamically created modules. We use a cluster search engine to deliver many of our pages including Topic pages. So for the example Icing, where we can expose images of Rooms that show the Icing technique, expose all the Supplies associated with Icing, Types of Icing, and finally other related Techniques. Now we assume that establishing for example a Action/Product relationship Icing > Ices and then having that module triggered by the relationship type would be a driver for establishing that relationship. The same would hold true for the Action/Property relationship.

3. Problem Statement

Assuming that the items don't have correct classification outside of the 'Specialty', ask ourself these questions –

Does the product need its own visibility to justify the new node?

Is the product a new variation that has a bright future so there will be followers and more of that particular variation?

Is the item close enough to an existing specific node that it should be placed their and attributed/faceted out?

How much does that variation account for in the specialty node? When it starts reaching 10% of the product assortment in 'Specialty', the researchers decide to split them out.

Can you attribute out/facet the specialty nodes to help clarify what is there? Depending on your users, the facet/attribute approach often works quite well.

A couple of issues are to be discussed. We provide below an illustrative example as the background. Let us take a broad term, 'pumps'. One is to separate the taxonomy for 'pumps' from the taxonomy for the content/records that go with the pumps; two, consider going fractal – we assume as faceted on both, then intersect them as combinations. So, for the pump (which is a functional concept) we can propose the following questions:

pump - what (gas, water, oil, ginger, breast milk, concrete)
pump - action (reciprocating, suction)
pump - states (submersible, in-line, three-phase, heavy-duty, corrosion resistant)
pump - who (names of competitors, suppliers, installers, if applicable)
pump - discipline (Oil and Gas, Agriculture, Mining, Medicine, Sports)

Content and records are normally classified by function (document type/record type) alone, but we could easily add others. So, in addition to specification, manual, catalogue and bulletin we could add:

document - states (hard-copy, digital, on-line, archived, recent)
document - activity (service, maintain, repair, install)
document - event (publish, update, archive)

We can make a fairly extensive collection just by combining the values we have listed here, and we are certain that any text collection could be classified the same way.

No term could leave the good faceted classification system - sort of like accounting.

The interpretation of the above illustration is as follows: We actually do have a faceted classification system for classifying the products once they are in the product category of pumps (in addition to maintaining separate categories for the different types of pumps). An issue we run into though is for each of the facets (what, action, states, who, discipline, etc) we have an "Other" which ends up collecting a various assortment of things. So when do we take a look at that "other" bin for the "What" and say ok we need to add 5 more facets to this classification because there are now too many "others".

4. Proposed taxonomy of Web Content

The discussions above lead to laid down a structured architecture as below. For the proposed web page taxonomy, we include both the content-based and context-based features.

4.1. Selected Features

In the proposed method, the information of web page and the relations could be considered. In other words, the information of relations between possible types of Source as a virtual document is considered. To select and identify the relations between Source, we consider the two types of pages, viz., the Main Page, and Fractals. The Fig.1 shows these relations.

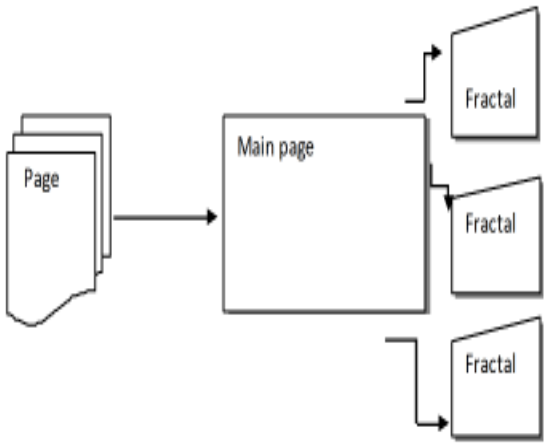

Figure 1. The relations between types of pages during taxonomy

The content relation is observed as below.

a) Title of s page which generate source data
b) Main content in page
c) Fractal in the related page

The source data of the web page is very important and simple to identify the main subject of a web page. In the same way the Anchor texts on the web page are usually around the main subjects of the page. By selecting the last feature, we go further in the relations' content to get some idea about the subject of the target page. In this way, we have considered both the content and the graph around the web page in our features selection.

4.2. Taxonomy inclusion

To show the usefulness of the combination of the selected features, we have exploited them in a taxonomy as follows. This proposed taxonomy is composed of two main modules called Graph Generation and Page Classifier. [6]. The applied dataset is the compute-related part of proposed source.

4.2.1. Graph Generation. In this module the graph around the target pages has been constructed. Fig.2 shows the generating process of the graph as our data set. We explain through the Figure 2, for generating the graph, there are two main components as follow:

a. Dataset parser

In the first step**,** according to the predefined Domain 1 of web page also the *Main* (our dataset), this component gathers the related pages with regarding the following.

1. We consider '*Main*' domain in this paper to 'Category' segment2 in each page, and then processes the page content to extract title, existing links (anchor text, URL) and category. Finally information or extracted metadata will be saved as a text file in metadata repository.

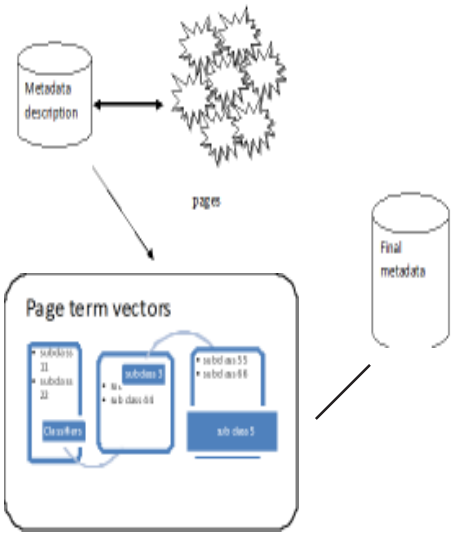

Figure 2. Web page graph for generating vectors

b. Metadata processor

This component takes the text files from metadata repository one by one and after extracting their links, a unique code is assigned to every URL. By using this information, it can construct the graph of dataset. This directed graph is represented as a matrix which its rows and columns imply URLs codes and their corresponding value (0 or 1) show the existence linkage between two pages.

5. Experimental Evaluation

This section describes a simple, direct method for evaluating semantic similarity, using human judgments as the basis for comparison.

We have proposed a method of classifying the Source web page documents by combining the Page term vectors and metadata. The results achieved with the current approach are quite encouraging. In most cases, the system was able to categorize each page in the most appropriate category. The few exceptions appeared due to limitations of the interpretation mainly because of the limitations in identifying the relation between main and fractal as the users have difficulties in expressing the relation using numbers.

One of our ideas for improving the efficiency of our Source classification system is using a suitable fractal selection method to optimize the combination of selected features. In addition, using some training sets, through a feedback mechanism, as we optimize the list of main and fractals for each class. Structure-oriented Weighting Technique also can be refined for obtaining better representations; however we refrain from using it as it depends different set of analyses.

6. Results

Table 1 summarizes the experimental results, giving the correlation between the similarity. Ratings (between main and fractals) and the mean obtained from experiments. We would like to stress that the fractal terms obtained from glossaries and taxonomies lead to limited number of pairs. Thus, the relation identification between main and fractals is not a complete detection of semantic relation for all the tested items. The similarity ratings by item are given in Table 3.

Similarity method Correlation	
Human judgments (for fractal)	r = 0.8435
Information content	r = 0.6430
Probability	r = 0.8368
Fractal relation counting	r = 0.8635

Table 1. The data obtained in the experimental results

	Page Term Vectors	**Sub class 11**	**Sub class 22**	**Sub class 33**
Automobile	3.92	3.424	3 .032	0.9962
Gem jewel	3.84	3.514	3.9286	3.0000
Travel	3.84	3.5 6	3.7537	3.9907
Ladder	3.76	3.5 8	3.4240	3.971
Ocean waves	3.70	3.5 1	3.8076	3.194
Wood house	3.61	3.6 15	3 .6656	3.291
Games wizard	3.50	3.5 1	3.6656	3.9999
Day weather	3.42	3.6 1	2.3925	3.9998
Fire stove	3.11	2.6 1.	1.7135 23	0.6951
Fruits basket	3.08	2.1 5	1.6 27	0.9689
Bird cock	3.05	2.2 9	.3139 29	0.9984
Bird crane	2.97	2.1 9	.3139 27	0.9984
Machine tools	2.95	3.4 6	.0787 29	0.9852
Monkey trees	2.82	2.4 2	.9683 24	0.8722
Crane implement	1.68	0.3 2	.9683 24	0.8722
Ladder steel	1.66	1.2	2.9355 26	0.8693
Travel car	1.16	0.7	0.0000 0	0.0000
Computer oracle	1.10	0.8	2.9683 24	0.8722
Food rooster	0.89	1.1	1.0105 18	0.5036
Hill area	0.87	0.7	6.2344 26	0.9867
Forest graveyard	0.84	0.6	0.0000 0 0	0.0000
Buddist monks	0.55	0.7	2.9683 27	0.8722
Tropical forest	0.42	0.6	0.0000 0	0.0000
Wooden furniture	0.42	0.7	2.9683 26	0.8722
Card reader	0.13	0.1	2.3544 20	0.8044
Glass painting	0.11	0.1	1.0105 22	0.5036
Noon meal	0.08	0.0	0.0000 0	0.0000
Ship travel	0.08	0.0	0.0000 0	0.0000

Table 2. Semantic similarity by item

Information-Based Semantic Similarity

n1 n2 main	(n1,n2) fractals
	fructose - sugar - 7.63 flowers
	fructose - sugar 3.56 - glucose
	fructose - glucose 8.26 – content

Table 3. Similarity with a sample test of relation among information content

7. Discussion

The experimental results in the previous section suggest that measuring semantic similarity using information content provides results that are better than the traditional method of simply counting the number for measuring the relations.

The measure also has a few difficulties as the count of main and fractals is user-dependent and sometimes lead to vast differences in measures. For example, Table 3 shows the word similarity for several words with fructose. Fructose and glucose are similar, both being sugar, and fructose and sugar are less similar, though not entirely dissimilar, since both can be classified as substances. The problem arises, however, in the similarity rating for fructose with flowers: the word flower is not the main or fractals for other terms, and as a result information-based similarity is maximized, and path length minimized, when the two words are both categorized as related ones. This is contrary to our intuition on measuring similarity and the taxonomies are approximation in reality.

Having considered a direct evaluation of the information-based semantic similarity measure, we discuss the possibility of the measure to address the syntactic ambiguity.

7.1 Coordination Ambiguity

Syntactic ambiguity is a chronic issue in natural language. As observed in [7], the class of "every way ambiguous" syntactic constructions | those for which the number of analyses is the number of binary trees over the terminal elements | includes such frequent constructions as prepositional phrases, coordination, and nominal compounds. In most of the research done in the information retrieval, researchers in natural language have made a great deal of progress in using quantitative information from text corpora to provide the needed constraints. Progress on broad-coverage prepositional phrase attachment ambiguity has been particularly notable, now that the dominant approach has shifted from structural strategies to quantitative analysis of lexical relationships (; Collins & Brooks, [8]. Noun compounds have received comparatively less attention [9] as has the problem of coordination ambiguity.

8. Conclusion

The process of word similarity detection is required in order to arrive at a correct interpretation of the information content. For example, analyzing main concept according to the structure of fractals could lead a similarity detection system to produce a noun phrase describing interpretation. Analyzing the main term according to the structure of the fractals could lead an information retrieval system to evolve consensus when looking for queries involving the term context.

The main and fractals relation establishment could be carried out with large metadata application.

References

1. McCallum, Sally H. (2004). An introduction to the Metadata Object Description Schema (MODS), *Library Hi Tech*, 22 (1) 82 – 88.

2. Glick, Gina M. (2005). METS. Metadata encoding and transmission Standard, Spring.www.courses.unt.edu

3. Kraft, Reiner ., Zien, Jason (2004). Mining anchor text for query refinement. International World Wide Web Conference archive., *In*: Proceedings of the 13th international conference on World Wide Web table of contents. New York, NY, USA. P. 666 – 674.

4. Resnik, Philip (1999) . Semantic similarity in a taxonomy: An Information based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*. 11. 95-130.

5. Zhang, S ., Bodenreider, O (2004). Comparing associative relationships among equivalent concepts across ontologies. Medinfo 2004, *In*: Proceedings of The 11th World Congress On Medical Informatics.p. 459.

6. Heymann, Paul., Garcia-Molina, Hector (2006). Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical Report. Stanford.

7. Church, Kenneth., Patil, Ramesh.(1982). Coping with syntactic ambiguity or how to put the block in the box on the table. Computational Linguistics. Volume 8 (3-4) 139 - 149.

8. Collins, M ., Brooks, J (1995). Prepositional phrase attachment through a backed-off model. Proceedings of the Third Workshop on WWW. 27-38.

9. Shirai, K., Tokunaga, T., Tanaka, H (1995). Automatic extraction of Japanese grammar from a bracketed corpus. Proceedings of the Fifth International Conference on Natural Language.