Exploring Lexicographic Ontologies for Hierarchically Organizing the Greek Wikipedia Articles

Maria Niarou[1], Sofia Stamou[2]
[1]Department of Archives and Library Science
Lonian University, 49100
Greece
[2]Computer Engineering and Informatics Department
Patras University, 26500
Greece
{niarou@gmail.com, stamou@ionio.gr}

Journal of Digital Information Management

ABSTRACT: *To effectively manage the proliferating online content, it is imperative that we come up with efficient data structuring and organization methods. Based on the findings of previous research [6] [7] that the most flexible and useful way to organize the online content is via the use of taxonomies and/or ontologies, we carried out the present study, which aims at structuring the content of the Greek Wikipedia via the use of the Greek WordNet. In particular, our study objective is to design a model that can automatically organize the Greek Wikipedia categories into a thematic taxonomy and based on the derived organization, to implicitly assign hierarchical structure to the encyclopedia articles that have been classified to the respective categories. To this end, we relied on the data encoded in Greek WordNet out of which we harvested the hierarchical relations that hold between the terms used to verbalize the Wikipedia categories. The effectiveness of our model is verified by the findings of several experimental evaluations conducted, which demonstrate that semantic networks are powerful resources for hierarchically organizing large volumes of dynamic data.*

Categories and Subject Descriptors
I.2.7 Natural Language Processing Text analysis: **H.3.1 Content Analysis and Indexing**

General Terms: Experimentation, Measurement, Performance, Human Factors

Keywords: WordNet, Wikipedia Articles, Hierarchical Organization

Received: 28 November 2011, Revised 4 February 2012, Accepted 1 March 2012

1. Introduction

Wikipedia is one of the most successful worldwide collaborative efforts to put together user-generated content, which could be used as an informational reference source for the online population. Currently, Wikipedia hosts millions of articles on a variety of topics, across different languages and has been incorporated into several computed-based applications. A crucial factor for its success is its open nature, which enables everyone edit, revise and/or question (via talk pages) the article contents. Considering the remarkable growth and the extensive use of Wikipedia, the question that rises naturally is: how can we assess the quality of Wikipedia or else how can we ensure that the content it provides is useful for its readers. In an attempt to shed light on the above issue, several researchers have proposed methods for assessing the quality of the Wikipedia articles and they have proposed methods for assisting Wikipedia editors provide qualitative and well-organized information in Wikipedia articles. Most of existing methods concentrate on the English Wikipedia (because it is the richest and the most widely used) although there have been successful attempts towards assessing and/or improving Wikipedia for other natural languages.

In this article, we study the structural quality of the Greek Wikipedia and specifically we investigate how we can organize the contents of the Greek Wikipedia so that we assist users experience successful navigations in its contents. Therefore our study is situated in the area of information management and combines tools and techniques from the field of natural language processing. In particular, we introduce a model which exploits the WordNet [8] semantic network for hierarchically organizing the Greek Wikipedia articles [I]. The motive for our study is to turn the Greek Wikipedia corpus into a structured data source and the reason for selecting WordNet as our reference guide for data structuring is the fact that it hierarchically organizes the concepts it contains based on their underlying semantic relations. The goal of our work is to experimentally demonstrate the contribution of

semantic networks into the hierarchical organization of online unstructured content. In this respect, we have designed and implemented a model that automatically captures the underlying semantic relationships that hold between the Wikipedia categories and based on their identified semantic links, it organizes them into a thematic hierarchy.

For unravelling the semantics of the Wikipedia categories as well as for deriving evidence about their relations, our model explores the information encoded in WordNet, a rich source of highly structured semantic information. The contribution of WordNet is mainly pronounced in the process of disambiguating the terms used to name the Wikipedia categories, as we will discuss later in the paper. In brief, our model operates on a three-step approach: firstly, it matches the Wikipedia category names to their corresponding WordNet nodes in order to extract their senses. Then, it disambiguates the categories matching several WordNet nodes based on their estimated semantic similarity to other categories with which they co-occur in the Wikipedia articles. Having detected the semantics of each Wikipedia category, we borrow the hierarchical structure of the category names from WordNet and apply them for organizing the categories into thematic hierarchies. Based on the above steps, our model automatically assigns the Wikipedia categories into hierarchical structures and as such it facilitates the organization of the Wikipedia articles that have been classified to the corresponding categories. The experimental evaluation of our model indicates that WordNet is a valuable source for semantically organising unstructured thematic data.

The remainder of the paper is organized as follows. We begin our discussion with a brief overview of relevant works. In Section 3, we introduce our model for hierarchically organizing the Wikipedia content via the use of semantic networks. In Section 4 we present the experiments we carried out in order to evaluate the effectiveness of our model and we discuss obtained results. In Section 5 we conclude the paper and we highlight our plans for future work. Before delving into the details of our work and how it differs from existing studies, we briefly present the resources deployed by our method, namely Wikipedia and WordNet semantic network.

2. Related Work

Numerous approaches have been proposed to hierarchically organize the Wikipedia articles using ontologies. [2] introduce Faceted Wikipedia Search, an alternative search interface for Wikipedia, which facilitates infobox data in order to enable users ask complex questions against Wikipedia's encoded knowledge. This way, users are allowed to query Wikipedia like a structured database and Faceted Wikipedia Search helps them to truly exploit Wikipedia's collective intelligence. In the same framework [19] describe how the BBC is working to integrate data and linking documents across BBC domains. In this case, DBpedia[1] is used as a controlled vocabulary to interlink the different thematic data units of BBC such as music, news, series etc.

In a different approach [16], the YAGO ontology has been applied for structuring Wikipedia. YAGO exploits the clean taxonomy of concepts from WordNet and annotates Wikipedia articles with conceptual labels. The annotation model is able to express entities, facts, relations between facts and properties of relations. By the YAGO model, all objects (e.g. cities, people) are represented as entities which can be linked in a relation. For example, in order to state that Albert Einstein won the Nobel price, we say that the entity Albert Einstein stands in the HASWONPRIZE relation with the entity Nobel Prize.

Among the most notable studies ranks also the development of Kylin Ontology Generator-KOG [19]. This system builds a rich ontology by combining Wikipedia infoboxes with WordNet and as a result integrates and extends the information provided by both Wikipedia and WordNet. KOG uses joint inference to predict subsumption relationships between infobox classes while simultaneously mapping the classes to WordNet nodes. KOG also maps attributes between related classes, allowing property inheritance. An ontology like this can greatly increase the recall of queries such as "What scientists born before 1920 won the Nobel prize?" by supporting transitivity and other types of inference. One of the major challenges is to convert the information present on the Web into Ontologies for the Semantic Web. Another aspect that is under-addressed and strictly related to the tasks of browsing and searching Wikipedia, is the similarity of articles at the semantic level. Given the previous remarks, [3] adopted a hierarchy of concepts (ontology) and a thesaurus to exploit links and provide a better characterization of Web data. To this end, they devised a system called THESUS (THEmatic SUsets of the WWW). The system deals with initial sets of Web documents, extracts keywords from all pages' incoming links and converts them to semantics by mapping them to a domain ontology. Then a clustering algorithm is applied to discover groups of Web documents.

Ontologies are also used in techniques for the automated hierarchal text clustering. Ideally, the hierarchical organization gives the opportunity to create object clusters, able to exploit the relationships between the object categories. A series of studies have been carried out aiming the creation of clusters [21] [22] [23], but without being freed from the human intervention. To deal with this problem, a recent study [24] attempts to develop a tool which allows the reduction of the human factor through three steps. According to this method, the procedure for the localisation of the wrong documents is facilitated by a good data clustering assumption.

[1]A community effort to extract structured information from Wikipedia and make this information available over the Web [IV].

It is also worth to make a reference to another study, which is occupied with the semi-automated web taxonomy construction [1]. The challenge, here, is to create high-quality links between the taxonomy topics. These links are supposed to be controlled and commented from the specialist of each knowledge field. Therefore, this method aims to facilitate the taxonomy specialists and not to replace them.

[12] in their study support that they have developed a solution, OntoGenie, that parses the Web pages to create knowledge instances for a given Ontology using WordNet as a bridge, mapping between Ontologies and the Web page terms. Thus, Ontologies[2] provide the explicit formalization and specification of the concepts and their corresponding relationships. This tool (OntoGenie) uses WordNet to convert unstructured data from the Web into structured knowledge for the Semantic Web. OntoGenie is a semi-automatic tool that takes as input domain ontologies and unstructured data from the Web and generates Ontology Instances (OI) for the given data, according to the above mentioned procedure.

Apart from the above, semantic networks and their hierarchical structure are also useful in the field of the efficient data extraction and retrieval. In [29] WordNet is used as a comprehensive semantic lexicon, in order to retrieve texts for the improvement of the communication, where the queries are expanded through the design of key-words. In this way, WordNet is used as a linguistic knowledge tool for the representation and the interpretation of terms, while it provides the user with a more efficient and integrated access to the information. In the same direction, [28] suggests the use of WordNet as a tool for the automated thesaurus drafting.

Another contemporary trend is the development of the semantic web. It is about a new form of the Web, "understandable" to the computers. The goal, here, is to provide wider functionality through the use of smart tools, such as information extractors. Practically, this new model is based on the comprehension of the word meanings, identifying important name entities like persons, organisations etc. via WordNet [33].

WordNet's contribution is also met in the field of the concepts' identification in a natural language. In this case, the semantic network plays a fundamental role during the processing of the word disambiguation, facing the problem of polysemy and synonymy, present in any natural language [34]. In the same time, researchers try to develop new disambiguation models in combination with the WordNet, producing ontological data bases through a systematic ordering of the polysemous terms contained in WordNet [26] or producing wide text bodies aiming to express words based on the relation of the synonymy in WordNet [27].

[2] Representation of domain knowledge in Semantic Web

It would be a default not to make a brief reference to some other applications of WordNet, such as the parameterisation of information systems, the teaching of a language and the machine text translation [30] [31] [32].

Despite existing works on thematically organizing Web data and the English Wikipedia, to the best of our knowledge no study has been reported about how to hierarchically organize the contents of the Greek Wikipedia in an automated manner. Too fill this void, we carried out the present study in which we propose a complete approach for turning the Greek Wikipedia into a thematic repository of structured knowledge. The details of our proposed method are given below.

3. Hierachical Organization of Wikipedia Articles Via Semantic Networks

Given the plentiful articles of Wikipedia, it is imperative that we deploy automated methods and tools for their hierarchical organization. To achieve a minimum level of data organization, the Wikipedia editors have attempted to organize rudimentarily the thematic categories under which the articles have been classified. However, this organization doesn't follow a clearly defined and scientifically documented structure. As a result, the upper regions of the Wikipedia categorization are almost exclusively thematic and no subsumption relation can be found among them, while the remaining reside unstructured inside the category network. Obviously, such categorization system is merely a partially organized thesaurus [13].

Contrariwise, WordNet contains very little domain-oriented knowledge, but at the same time it provides a well-structured top-level taxonomy. For that, we picked WordNet as our source for the hierarchical organization of the Wikipedia categories.

The goal of our approach is to exploit the structural advantages that each of the resources (i.e., WordNet and Wikipedia) has, in order to integrate them into a common thematic resource. In this regard we work on the present study, being largely inspired by the work of [13].

In the following sections, we present the proposed methodology and give a detailed description of the individual steps it entails, by providing documentation for every step, the weaknesses of our proposed method and we experimentally validate it. Then, we refer to the advantages and some points that need further investigation.

3.1 Methodology

In this section, we present in details the proposed methodology for the hierarchical organization of the Greek Wikipedia articles via the use of semantic networks. In brief, our proposed approach operates in the execution of the following steps.

1. Lexical processing and subsequent lemmatization of the Wikipedia categories, which are afterwards mapped to the WordNet synsets

2. Hierarchical graph structuring

3. Graph filtering

4. Taxonomy fine-tuning

To be more clear, the word tokens of every category name are lemmatized so that they can be mapped to their corresponding WordNet synsets. During this process, the mapping concepts of the Wikipedia categories adopt the corresponding senses from WordNet and their corresponding semantic relations. As a result, the Wikipedia categories are organized in a semantic hierarchy according to WordNet structure and Wikipedia is ontologized. In the following subsections, we describe extensively the steps of our methodology, documenting our design decisions and giving reasons for their necessity.

3.1.1 Mapping

In the first step, we take as input the Wikipedia categories and parse their literals in order to be mapped against their corresponding WordNet synsets. This is an important and complex task because, unlike WordNet synset names which are comprised of lemmatized terms, a large fraction of the Wikipedia categories have complex names expressed via inflected wordforms and some of which contain also stopwords. This lexical complexity in verbalizing Wikipedia categories is largely due to the fact that their names are manually determined and reflect human intuition about conceptual organization of the world knowledge [10]. Accounting that Wikipedia category names do not correspond to the type of lexical concepts we would expect to encounter in texts, it becomes more than evident that we need to process them in order to transcribe them into forms that are compatible with WordNet synsets' naming. This pre-processing of Wikipedia categories concerns their Part-of-Speech tagging and subsequent lemmatization. For this task, we extract the Wikipedia category names and give them as input to the Brill Part-of-Speech tagger [1], which assign to every category word token an appropriate grammatical category and identifies the first inflected form (i.e., lemma) for every term.

Thereafter, based on [15], we parse the lemmatized Wikipedia category names in order to identify those that correspond to named entities and detect the types of the latter. Having discriminated between proper and common named category entities, we further process the latter by applying shallow syntactic parsing to their word constituents. Here we should note that, parsing of the category names concerns only multi-term categories for the names of which there exists contextual information upon which we would identify their syntactic roles. For parsing multi-term Wikipedia category names, we rely on the Treebank syntactic dependency parser [18], which we largely employ for detecting the role of stopwords (i.e., prepositions, conjunctions, etc.) within category names. Focusing on the functional role of stopwords, such as prepositions, enables us to detect latent temporal (e.g., "during"), spatial (e.g., "in"), causal (e.g., "by"), etc. orientations of the category names. Such orientations will be further utilized in the evaluation process described in the next section.

Now, back to the processing of the Wikipedia categories, we rely on the parsed category names, the tokens of which we look up against the WordNet synset names. We remind here, that all word tokens in category names are lemmatized and therefore they can be successfully mapped to their corresponding WordNet synsets, unless the latter are absent from the network.

3.1.2 Graph structuring

At the end of the first step, we have the lemmatized Wikipedia categories mapped to suitable WordNet synsets. In the second step, we look for semantic relations between Wikipedia categories, based on the relations between the respective concepts encoded in WordNet. For example, for every Wikipedia category we rely on all matching WordNet synsets (in case of polysemous category names) and for every matching synset we examine if it is linked in WordNet to any other Wikipedia category via an IS-A relation. This is performed by traversing – starting from the Wikipedia matching synset- the WordNet hierarchy upwards, until: i) a Wikipedia category name is found, or ii) a WordNet root node is encountered. Following these steps, one or more conceptual trees are produced, which link together hierarchically the examined Wikipedia categories according to their respective hierarchical structure in WordNet. Note that, the number of the emerging conceptual trees is determined by the polysemy degree (i.e., number of matching synsets) of the examined Wikipedia category as well as the presence of semantic relations (in WordNet) between the considered Wikipedia categories. In case an examined Wikipedia category is disjoint in WordNet with the remaining Wikipedia categories, it is omitted from the deduced conceptual trees. Having built the trees, our next step is to weight their edges. We remind that all tree edges represent IS-A relation types, given that our objective is to organize Wikipedia categories into a thematic hierarchy. The weighting of the edges is necessary in order to retain every Wikipedia category, belonging to several trees, into a single hierarchy (the more representative one). Weighting of the edges corresponds into assigning a score to every link in each tree, in order to indicate the closeness (i.e., proximity) between the linked concepts in both Wikipedia and WordNet. Proximity between two nodes (each node representing one concept) of the graph is expressed as the inverse distance of paired nodes in the graph, formally given by:

$$Proximity\,(v_i, v_j) = 2\,\frac{1}{D_{WordNet}\,(v_i, v_j) + D_{Wikipedia}\,(v_i, v_j)}$$

Where (v_i, v_j) is the examined pair of nodes (i.e., the pair of concepts for which the proximity is examined),

$D_{WordNet}(v_i, v_j)$ represents the distance of the respective nodes (v_i, v_j) in WordNet and $D_{Wikipedia}(v_i, v_j)$ represents the distance of these nodes (v_i, v_j) in Wikipedia[3].

3.1.3 Graph filtering

In case the name of a Wikipedia category belongs to more than one conceptual tree - due to polysemy – it is necessary to select the tree which is more appropriate to host the respective category. By doing so, every Wikipedia category will belong to a single hierarchical tree. The category matching nodes which appear in multiple trees can be filtered based on their Proximity score, in the sense that a category concept is retained to the hierarchy for which it has the highest proximity value to its related category concepts. As a result, the names of the Wikipedia categories are disambiguated, as each of these takes the sense of the WordNet concept to which it is mapped and which is the most appropriate according to the estimated proximity value.

3.1.4 Taxonomy fine-tuning

Given that, at the end of the previous step, some of the Wikipedia categories may be left unstructured, we handle them as follows. We build a secondary list of topics where we store these Wikipedia categories associated with their corresponding articles. The scope of the secondary index is to be used as a complementary unstructured list of topics, helpful for information research and understanding the content of the respective articles.

At the following table, we resume some statistical information concerning the number of the Greek Wikipedia categories and the fraction of which are joined in the conceptual hierarchies, after the implementation of our methodology. Note that this information concerns the content of the Greek version of Wikipedia during October 2010.

Greek Wikipedia Categories	6,158
Inter-linked categories in Greek Wikipedia	985
Avg. links/ category	1.5
Inter-linked categories according to WordNet	4,293
New hierarchies	825
Avg. nodes/ hierarchy	5.2
Unlinked categories	880

Table 1. Statistical information about the linking of Wikipedia categories

As the table shows, Greek Wikipedia contains 6,158 categories of which only 985 are inter-linked to other categories with an average number of links per category name of 1,5. The low fraction of linked categories further supports the motive for our work, which is to come up with automatic tools for organizing Wikipedia categories. This is attested in the fact that, after applying our method to the Greek Wikipedia, 4,293 more Wikipedia categories are linked with other. This is because the concepts which describe them are found in WordNet and demonstrate semantic relations to other category concepts. Thus, the percentage of linked categories is increased by 69,71%, which means that, after the implementation of our methodology, the total percentage of the linked categories reaches up to 85,70%. Concluding, 880 Wikipedia categories remain disjoint, even after implementing our methodology, because the concepts which verbalize them are absent from the WordNet network.

The following figure illustrates an example of how the Greek Wikipedia categories are organized into a hierarchy after the application of our method.

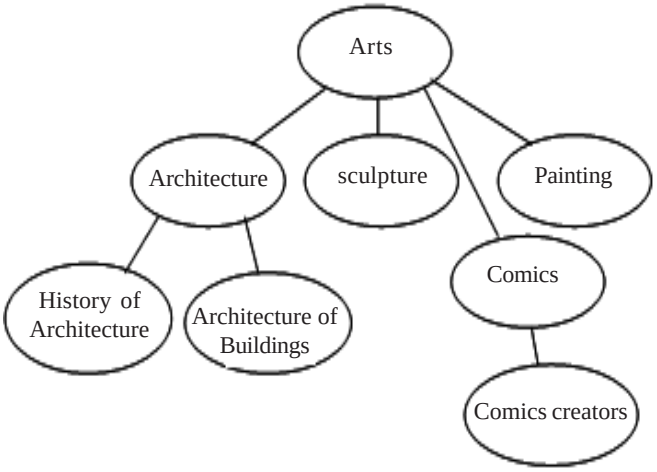

Figure 1. Example of the hierarchical structure assigned by our method to a sample of the Greek Wikipedia categories

Last but not least, we present the algorithmic steps that our method takes for organizing the Greek Wikipedia categories into topical hierarchies.

Input: set of Wikipedia categories **C**, **W**ord**N**et, **PoS**-tagger, **Parser**, **N**amed **E**ntities recognizer
Output: hierarchically structured WC
For each category c in C do
Extract all terms T describing the name of c
PoS tag every term $t_{(i)}$ in T
Associate every $t_{(i)}$ with corresponding lemma, $lem_{(i)}$
Return lemmatized category names for categories C
end
For each $lem_{(i)}$ in C run Named Entities recognizer
if $lem_{(i)}$ is NE
Return entity type and annotate $lem_{(i)}$
else

[3] The values of the proximity are standardized [14] and take values between 1 (indicating very strong semantic proximity, i.e. direct hyponymic relation) and slightly above 0 (indicating very weak semantic proximity).

Apply syntactic parsing
Return syntactic tags and annotate $lem_{(i)}$
end
end
For each syntactically annotated $lem_{(i)}$ in C
Map $lem_{(i)}$ to WordNet nodes
if matching found
Return sense of matching node
else
Store $lem_{(i)}$ in secondary index
end
end
For each WN matching sense for $lem_{(i)}$ do
Extract WN and Wikipedia semantic structure for $lem_{(i)}$
Compute Proximity for every sense of $lem_{(i)}$ with every other lem in C
Sort sense of $lem_{(i)}$ by ascending Proximity scores
if proximity scores identical for all senses of $lem_{(i)}$
Select sense of most dense structure
Return structure of selected sense
else
Select sense of max Proximity
Return structure of selected sense
end
end

The above process is also schematically illustrated in the following figure.

3.2 Open issues on hierarchical organization

Having described in details the proposed methodology, we should point out a number of open issues in relation to the hierarchical organization of Wikipedia categories. In particular, as we have already mentioned, our methodology does not reject the links between Wikipedia categories which are absent from WordNet. Instead, we choose to retain the semi-structured Wikipedia categories in secondary thematic indexes, either as a semi-structured list of topic, or as a non-structured file, so as to ensure that article readers will have access to the topical content of the articles. In the future, filtering techniques for the disjoint Wikipedia categories could be studied. The reason why we don't attempt this now is that, for this paper, we use the Greek version of both Wikipedia and WordNet, which lack in wealth from their English counterpart. Indicatively, we note that the English version of Wikipedia includes 3,891,085 articles, whereas the Greek Wikipediacontains only 70,500 and the English WordNet contains nearly 150,000 synsets when the Greek one includes approximately 22,000 synsets[4]. The only way to enrich the Greek version of these two resources is to increase the number of volunteers who edit them and who would be able to adopt translation techniques in case of semantic equivalence of the contents between the two language sources (Greek and English). Therefore, any comparison between the English Wikipedia and WordNet to their Greek counterparts would not deliver any useful results as there is no direct semantic or pragmatic equivalence between the respective resources. In addition and with respect to the Greek language sources, there are only a few attempts to evaluate their quality and as such there is not sufficient evidence about their contents organization and/or structural properties. Consequently, it is more than clear that the content or the structure of these resources cannot be directly compared to their English equivants. Especially because there is no content correlation neither between the Greek and English version of Wikipedia nor between these of WordNet. Additionally, we don't aim to a comparative evaluation of the two resources or to improve the structure of the English version of Wikipedia. Our goal is to automatically organize efficiently the content of the Greek version of Wikipedia based on the exploitation of the English version of it, the WordNet and some techniques relevant to them which have been used and incorporated to Wikipedia.

3.3 Contribution

The derived hierarchies of Wikipedia categories can be used in various applications. Given that, the average degree of polysemy in Wikipedia categories accounts to 3.2 senses per category name, it becomes evident that disambiguation of the category semantics constitutes an indispensable step towards taxonomy structure extraction for the manually selected Wikipedia categories.

Another advantage of our methodology is its independence from the information resources and the language in which they are written. This is because semantic networks such as WordNet, online encyclopedias such as Wikipedia, thematic taxonomies such as web directories etc., are available for most natural languages. Consequently, the exploitation of these resources for the structured organization of manually created information resources is more than feasible.

A third advantage of our approach is the fact that it is based on the Wikipedia categories and not on the article contents per se. In this way, we can minimize the necessity of frequently updating the derived hierarchies, given that the categories of Wikipedia are modified much rarely than the respective articles (except "feature articles"). Moreover, the Wikipedia categories are considered to be completed, as they cover a wide range of topics which can classify the human knowledge. This can be observed by the fact that 3,487 distinct Wikipedia categories constitute root nodes.

Concluding, one more advantage of our methodology is the automated mode that it works; which means that it is activated by the availability of a body text and there is no need for annotated examples to produce a model of learning the schema of hierarchical organization of categories.

[4]These numbers correspond to the period of March 2012

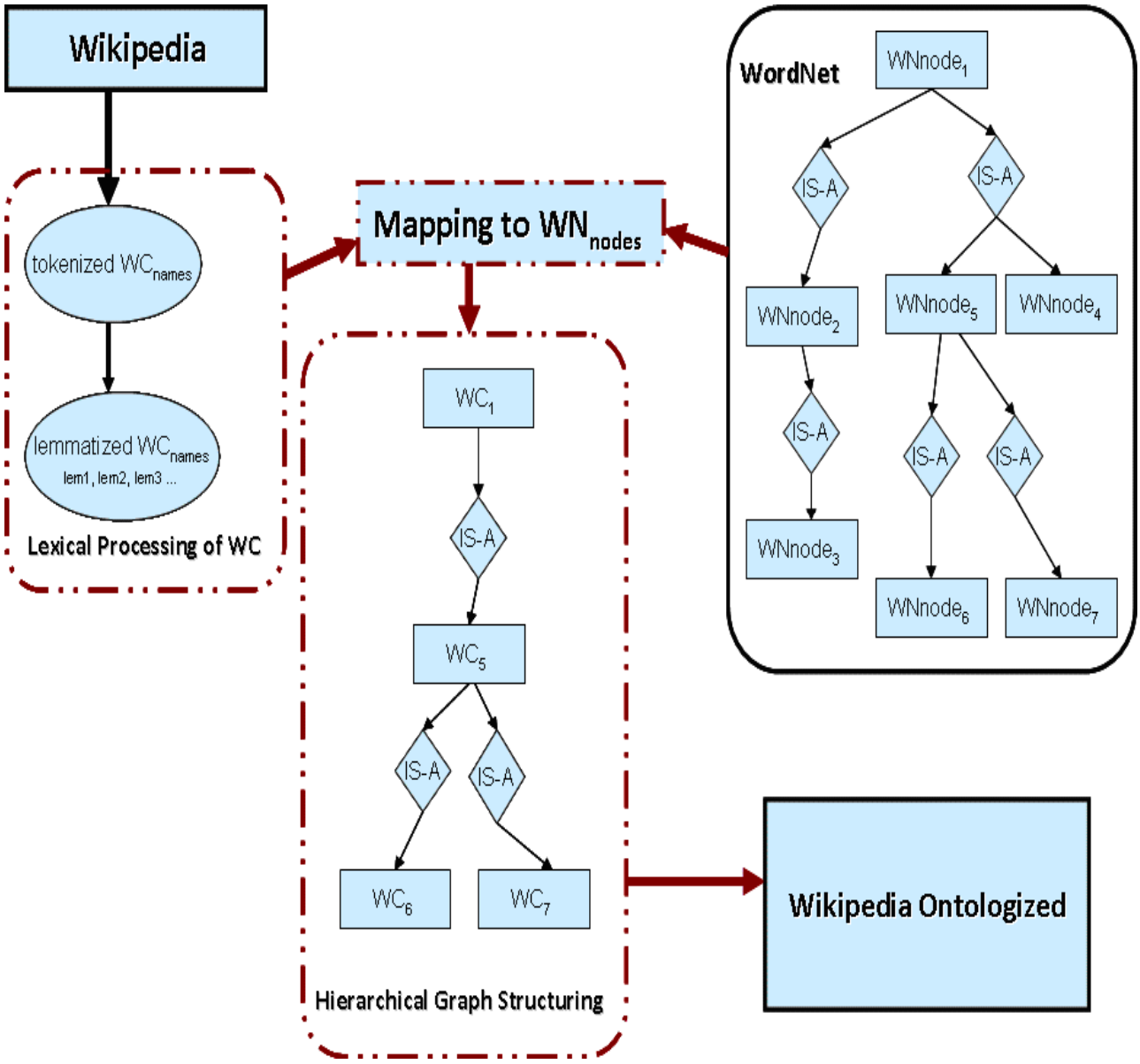

Figure 2. The algorithm's workflow

4. Experimental Evaluation

In this section, we present the experimental evaluation of our model for hierarchically organizing the Wikipedia content via the use of semantic networks. Due to the absence of a standard benchmark for evaluating the usefulness of the semantic networks in hierarchically content organization, we carried out two individual and complementary evaluations; user-centric and automated. In the following sub-sections, we present the experiments we carried out in order to evaluate the effectiveness of our model and we discuss obtained results.

4.1 User-Centric Evaluation

We carried out a user-centric evaluation, in order to obtain some user-related outcomes. User-centric evaluation concerns a process, which imposes the direct participation of the users. We carried this evaluation out aiming at thequality assessment of the derived hierarchical organization of the Wikipedia categories.

The participants in our survey are common users of Wikipedia, who worked according to the given guidelines. Specifically, we recruited 10 volunteers, 5 male and 5 female, between 22 and 28 years of age. All our participants are postgraduate students in Computer Science Departments of Greek Universities.

To conduct our study, we relied on the links our method delivered for connecting the Greek Wikipedia categories, delivered to our participants the 4,293 categories - linked according to WordNet - and asked them to evaluate as many links as they could during a one week period. This evaluation touches upon whether the derived links are correct and comprehensive in terms of user judgments.

After having collected the user feedback and in order to interpret it, we selected the links evaluated by at least 5 volunteers and implemented the metrics of Correctness and Comprehensiveness, explained below. Correctness is formally given by:

$$Correctness\,(r_i, (c_j, c_k)) = \frac{|Positive\ votes\ for\ (r_i,(c_j, c_k))|}{|total\ votes\ for\ (r_i, (c_j, c_k))|}$$

And evaluates whether a relation (r_i) which connects a pair of Wikipedia categories (c_j, c_k) is correct. Given the formula above, the greater the value of the metric, the more powerful is the degree of correctness of the examined pair of categories.

In a similar manner we assess the Comprehensiveness of the relation (*ri*) between a pair of Wikipedia categories (c_j, c_k) as:

$$Comprehensiveness\,(r_i, (c_j, c_k)) = Correctness\,(r_i, (c_j, c_k)) \bullet \frac{V(r_i,(c_j, c_k))}{V}$$

Where, $V(ri,(cj, ck))$ is the number of the participants who evaluated a given relation (ri) for the categories (cj,ck) and V is the number of the participants who evaluated all the examined relationships. According to the above formula, the greater the value of the metric, concerning the comprehensiveness of a relation between two categories, the more powerful is the convergence of the evaluators on the correctness of the examined linking.

Based on the above metrics, we estimate how accurate and understandable are the links, which our method identified, between the Wikipedia categories. In other words, we assess whether the given hierarchical relationships between pairs of Wikipedia categories are correct and comprehensive.

We conducted the user-centric evaluation in order to assess the contribution of the given algorithm in the hierarchical organization of the Wikipedia categories. Consequently, we can also assess the ability that our methodology provides towards the organization and management of the Wikipedia content associated with the examined categories. Table 2 reports the fraction of correct and comprehensive links our algorithm identified. Note that the reported results pertain only to links assessed by all participants.

According to the results, the total number of the evaluated links comes to 821. From these, we take into account only the results concerning the 280 links of Wikipedia categories, which were examined by at least 5 volunteers. We don't do any further examination to the judgments made by less than half of our subjects in order to avoid obtaining partial results. According to obtained results, 236 of the 280 links have been characterized as correct and only 44 have been characterized as incorrect. In other words, 84.28% of the links between the Wikipedia categories that were examined by at least 5 volunteers have been characterized as correct. This finding verifies the capacity of our method towards organizing the Wikipedia categories hierarchically. Moreover, it is noticeable that the respective average of the linked categories' comprehensiveness reaches up to 76%, which indicates that human judgments on the links' accuracy are similar enough to each other; thus indicating increased levels of annotators' agreement.

The main advantage of the user-centric evaluation is the direct participation of end user in the evaluation process; a fact that helps us capture the usefulness of the derived hierarchies. Besides, it gives a clear picture of the way users perceive the derived thematic hierarchies.

The completeness of this type of evaluation depends on its duration, on the number of participants involved and of course on the richness of the derived hierarchies under evaluation. But, considering the amount of different thematic Wikipedia categories and the number of senses that WordNet encodes for each of the categories, it becomes obvious that user-centric evaluation can't be totally completed as it is difficult to evaluate one by one all links between Wikipedia categories. On top of that, we should not ignore that user-centric evaluation entails a level of subjectivity as it explores the participants' perception of the hierarchies' correctness and usability. Consequently, there are inherent weaknesses and limitations. To tackle the disadvantages of this evaluation study, we carried out one more evaluation experiment in which we automatically assessed the contribution of our proposed approach towards structuring the Wikipedia categories. The details of the automatic evaluation are presented next.

4.2 Automated Evaluation

We carried automated evaluation out in complementation with the user-centric and in order to minimize the

Evaluators	10
Evaluated hierarchical relations	280
Positively evaluated hierarchical relations	236
Negatively evaluated hierarchical relations	44
Avg. Correctness of relations	0.8428 (84.25% correct)
Avg. Comprehensiveness of relations	0.76 (76% comprehensive)

Table 2. Statistical information from the user-centric evaluation

percentage of subjectivity in our evaluation results. This type of evaluation is based on an automated process. Our goal is to examine the effect of the derived hierarchical structure on the performance of automatic techniques for organizing and managing data.

Two statements constitute the basis of our study:

- Each article in Wikipedia is assigned to one or more thematic categories by its editor or/and by its administrators
- The majority of the Wikipedia articles are connected by internal links, which indirectly indicate factual relationship between the linked articles. Thus, internal links can be considered as indicators of thematic relevance between two texts.

Based on the above admissions, we examine the categories to which the linked (via internal links) articles are assigned. If these categories are also connected at the derived hierarchy, we can assume that our methodology is useful for selecting related texts and consequently the providing structure of the categories is correct. In contrary, if the categories of articles which are not linked in Wikipedia, are presented connected in the derived hierarchy or the reversed, we assume that our methodology isn't useful for selecting related texts and consequently, the providing structure of the categories isn't correct.

To carry this study out, we select a subset of Wikipedia articles linked in the derived hierarchy and we take it as input for the direct evaluation of the correctness and the indirect evaluation of the comprehensiveness of this hierarchy. We extract the thematic categories of these articles from the Wikipedia dump file and we map these categories to the nodes of the derived hierarchy. Then, we calculate their relevance degree based on the metric *OSim* [4] explained below.

To be more specific, we selected some Wikipedia categories which describe a random sample of 100 Wikipedia articles connected via internal links. Given the connection between these articles, we can suppose that there would be a corresponding relation between the categories to which the articles are assigned. In particular, the selected sample of 100 articles is assigned to 314 Wikipedia categories. To ascribe relations to these 314 Wikipedia categories, we based on the corresponding links according to their articles and we concluded to a sample of links for the evaluation of the performance of our algorithm. More detailed, we examined the way that these categories are linked in the derived hierarchy and we compared their overlap with the link at the encyclopedia. As we have already mentioned, for the calculation of the overlapped relations we implement the metric *OSim*.

Part of our work, is also the implementation of the present evaluation formula in order to evaluate the contribution of our methodology at the hierarchical organization of Wikipedia categories. At the same time and in this way, we can also evaluate whether our methodology gives us the ability to organize and manage the encyclopedia content assigned to these categories.

The formula of the metric *OSim* is:

$$OSim\,(C_A, C_B) = \frac{(C_A \cap C_B)}{C}$$

Where C_A represents the categories of the articles linked via internal links, C_B represents the categories connected in the derived hierarchy and *C* represents the total of the examined categories. The metric above, calculates the overlap degree of the links between the examined categories. The more categories of linked -via internal links- articles are presented connected at the derived hierarchy, the best the organization of these categories is and consequently, our methodology is effective and the opposite.

Implementing the given evaluation formula and as it is indicated in the table below (Table 3) we observe that the number of the Wikipedia categories linked in the encyclopaedia is 314 whereas the number of categories linked in the derived hierarchy is 268. In other words, the overlap of the relations between the Wikipedia categories at the extent of the derived hierarchy and the Wikipedia is 85.35%. This element strengthens our previous claims referring on the effectiveness of our algorithm for giving hierarchical structure at the encyclopedia data. In particular, Table 3 illustrates that for the 100 Wikipedia articles that have some connection to each other and which we sampled for our automatic evaluation, the total number of their inter-connected categories amounts to 314. Note that those 314 article categories have been manually assigned their internal links as determined by the article editors. After applying our Wikipedia category structuring method to the same set of articles and corresponding categories, the number of inter-linked categories we obtained amounts to 268. To be able to interpret obtained results we applied to the OSim formula and we examined the amount of overlapping elements in the two sets of interlined Wikipedia categories (one set of the manually linked categories and one set of the automatically linked categories). As the Table shows the overlapping links between the examined categories reached to 85%, this indicating that n 85% of the cases our method managed to establish as good connections between categories as humans would do.

Internally linked articles	100
Categories of linked article	314
Categories inter0linked in our hierarchy	268
OSim	0.8535
% of overlapping links between categories	85.35%

Table 3. Statistic information for the automated evaluation

The automated evaluation of the hierarchical organization

of the Wikipedia categories contributes in carrying out a large-scale evaluation without the cost of human resources, because it doesn't demand a user-centric process. Moreover, the results are objective and universally valid, as they don't rely on users' judgments. The automated evaluation also ensures that all the links between the thematic categories can be evaluated in immediate time. This allows the indication of possible mistakes or gaps and as a result their correction.

Concluding, the present type of evaluation can also be useful for the enrichment of WordNet semantic network by suggesting links and/on nodes missing and can be added.

5. Conclusions

In this paper, we study the hierarchical organization of the Greek Wikipedia articles using a semantic network. Our goal is to propose and implement a model against the hierarchical data organization. Initially, we presented the subject of our study and we made an introduction to it. Then and after a brief overview of the related work, we described the proposed methodology for hierarchically organizing the Wikipedia contents via the use of a semantic network. To design our methodology, we exploited the strong aspects of the two knowledge resources upon which we relied, namely Wikipedia and WordNet in order to unify them into a common resource so that Wikipedia is ontologized and its content is much more organized.

Additionally, we described two fundamental approaches to evaluate the derived hierarchy of the thematically structured Wikipedia categories. Experimental results indicate that semantic networks are powerful sources for assigning hierarchical structure to the topical concepts selected by Wikipedia editors for organizing the encyclopedia's content.

Based on the findings of our current work, we point out avenues for future research, some areas of which are listed below:

- exploit more available ontologies for the hierarchical organization of the Wikipedia categories
- provide advanced navigation services in the articles of Wikipedia
- support faceted search against Wikipedia content
- enrich the WordNet semantic network with senses derived by the Wikipedia thematic categories.

Overall, the reported study is the first attempt to bring hierarchical organization to the Greek Wikipedia, a resource that constantly gains popularity and visibility among online referential resources.

References

[1] Brill E. *A SIMPLE Rule-Based Part of Speech Tagger.* PhD Thesis. Department of Computer Science University of Pennsylvania Philadelphia, Pennsylvania

[2] Hahn, R., Bizer, C., Sahnwaldt, C., Herta, C., Robinson, S., Bürgle, M., Düwiger, H., Scheel, U. (2010). Faceted Wikipedia Search, *In:* Proceedings of the 13th International Conference on Business Information Systems *(BIS 2010)*, Berlin, Germany.

[3] Halkidi, M., Nguyen, B., Varlamis, I., Vazirgiannis, M., (2003). THESUS: Organizing web document collections based on link semantics, *Journal of Very Large Databases*.

[4] Haveliwala, T. (2002). Topic Sensitive PageRank. *In*: Proceedings of the 11th Intl. World Wide Web Conference, p. 517-526

[5] Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R. (2009). Media Meets Semantic Web: How the BBC Uses DBpedia and Linked Data to Make Connection. *In:* Proceedings of the 6th European Semantic Web Conference Research and Applications, p. 723-737, Heraklion, Crete, Greece.

[6] Liu, S., Liu, F., Yu, C., Meng, W. (2004). An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases. *In:* Proceedings of the International Conference on Information Retrieval.

[7] Mandala, R., Tokunaga, T., Tanaka, H., Akitoshi, O., Satoh, K. (1998). Ad Hoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri, *In: TREC*, 7, 414-419.

[8] Miller, G. A. (1995). WordNet: A Lexical Database for English, *Communications of the ACM.* 38 (11) 39-42.

[9] Miller, G. A., Fellbaum, C. (2007). WordNet then and now, *Language Resources and Evaluation*, 41, 209-214.

[10] Nastace, V., Strube, M. (2008). Decoding Wikipedia Categories for Knowledge Acquisition, *In*: Proceeding of the 23rd national conference on Artificial intelligence, Vol. 2

[11] Palmer, M. (2000). Consistent criteria for sense distinctions, *Computers and the Humanities,* 34, 217-222.

[12] Patel, C., Supekar, K., Lee, Y. (2005). OntoGenie: Extracting Ontology Instances from the World Wide Web. *In*: Lecture Notes in Computer Science, 3428, 51-63.

[13] Ponzetto, S., Navigli, R. (2009). Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. *In*: Proceedings of the 21st International Joint Conference on Artificial intelligence, Pasadena, California, USA, p. 2083-2088.

[14] Ramakrishnan, R., Gehrke, J. (2002). Database Management Systems p. 728-737.

[15] Stamou, S., Kozanidis, L. (2009). Towards Faceted Search for Named Entity Queries. *In*: Lecture Notes in Computer Science (LNCS), 5731, 100-112

[16] Stamou, S., Oflazer, K., Pala, K., Christodoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D. Grigoriadou, M. (2002). BALKANET: A Multilingual Semantic Network for Balkan Languages. *In:* Proceedings of the 1st International WordNet Conference, Jan. 21-25, Mysore, India.

[17] Suchanek, F., Kasneci, G., Weikum, G. YAGO: a Core of Semantic Knowledge Unifying WordNet and Wikipedia [available at: http://www2007.org/papers/paper391.pdf].

[18] Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., Nivre, J. (2008). The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. *In*: Proceedings of the 12th Conference on Computational Natural Language Learning, p.159-177.

[19] Wu, F., Weld, D. (2007). Autonomously Semantifying Wikipedia. *In*: Proceedings of the 16th International ACM Conference on Information and Knowledge Management, Lisbon, Portugal, p. 41-50.

[20] Wu, F., Weld, D. (2008). Automatically Refining the Wikipedia Infobox Ontology. *In*: Proceeding of the 17th International World Wide Web Conference. p. 635-644, China.

[21] Koller, M. (1997). Hierarchically Classifying Documents Using Very Few Words. *In:* Proceedings of the 14th International Conference on Machine Learning, p. 170-178.

[22] Ruiz-Casado, M., Srinivasan, P. (2002). Hierarchical Text Categorization Using Neural Networks. *In*: Information Retrieval, 5 (1) 87-118.

[23] Wang, K., Zhou, S., Liew, S. (1999). Building Hierarchical Classifiers Using Class Proximity. *In:* Proceedings of the 25th International Conference on Very Large Databases.

[24] Adami, G., Avesani, P., Sona, D. (2003). Clustering Documents in a Web Directory. *In:* Proceedings of the 5th ACM International Workshop on Web Information and Data Management, p. 66-73.

[25] Kumar, R., Raghavany, P., Rajagopalan, S., Tomkins, A. (2001). On Semi-Automated Web Ttaxonomy Construction. *In:* Proceedings of the International Workshop on Web and Databases, p. 91-96.

[26] Buitelaar, P. (1998). CORELEX: An Ontology of Systematic Polysemous Class. *In:* Proceedings of the FOIS Conference, p. 221-235, IOS Press, Amsterdam.

[27] Mihalcea, R., Moldovan, D. I. (1999). Automatic Acquisition of Sense Tagged Corpora. *In:* Proceedings of the 12th International Florida Artificial Intelligence Research Society Conference, p. 293-297.

[28] Mandala, R., Tokunaga, T., Tanaka, H., Akitoshi, O., Satoh, K. (1998). Ad Hoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri, *In:TREC*, 7, 414-419.

[29] Van de Riet, R., Burg, H., Dehne, F. (1998). Linguistic instruments in information system design. *In:* Proceedings of the 1st International FOIS Conference, IOS Press, Amsterdam.

[30] Chai, J. Y., Biermann, Á. W. (1999). The Use of Word Sense Disambiguation in an Information Extraction System. *In*: Proceedings 16th National Conference on Artificial Intelligence, p. 850-855, AAAI Press, Menlo Park (Ca).

[31] Hu, X., Graesser, A. (1998). Using WordNet and Latent Semantic Analysis to Evaluate the Conversational Contributions of Learners in Tutorial Dialogue. *In:* Proceedings of the ICCE Conference, p.337-341, China Higher Education Press, Beijing.

[32] Shei, C. C., Pain, H. (2000). An ESL Writer's Collocational Aid. *In*: Computer Assisted Language Learning, 13 (2) 167-182.

[33] Mihalcea, R., Mihalcea, S., (2001). Word Semantics for Information Retrieval: Moving One Step Closer to the Semantic Web. *In:* Proceedings 13th IEEE International Conference on Tools with Artificial Intelligence. ICTAI, p. 280-287.

[34] Moldovan, D. I., Mihalcea, R. (2000). Using WordNet and Lexical Operators to Improve Internet Searchers. *IEEE Internet Computing*, 4 (1) 34-43.

I. http://el.wikipedia.org/wiki/

II. http://www.illc.uva.nl/EuroWordNet/

III. http://blum.sabanciuniv.edu:8888/balkanet/

IV. http://wiki.dbpedia.org/About

V. http://en.wikipedia.org/wiki/Wikipedia:About