

# Wikipedia in action: Ontological Knowledge in Text Categorization

Maciej Janik

Large Scale Distributed Information  
Systems Lab (LSDIS)  
Department of Computer Science,  
University of Georgia  
410 Boyd Graduate Studies Research  
Center, Athens, GA 30602-7404  
janik@cs.uga.edu

Kryś Kochut

Large Scale Distributed Information  
Systems Lab (LSDIS)  
Department of Computer Science,  
University of Georgia  
410 Boyd Graduate Studies Research  
Center, Athens, GA 30602-7404  
kochut@cs.uga.edu

## ABSTRACT

We present a novel automatic text categorization method based on the knowledge represented in an ontology. An important feature of our approach is that it does not require a training set, which is necessary in the traditional supervised categorization methods, typically based on probabilistic approaches. In the presented method, the ontology, including the domain concepts organized into categories and interconnected by relationships, as well as instances and connections among them, effectively becomes the classifier. Our method focuses on (i) converting a text document into a thematic graph of entities occurring in the document, (ii) ontological classification of the entities in the graph, and (iii) determining the overall categorization of the thematic graph, and as a result, the document itself. In evaluating our categorization method, we used an RDF ontology created from the full English language version of Wikipedia. The presented experiments, conducted on corpora of news articles, showed that our training-less categorization method achieved a good overall accuracy.

## Categories and Subject Descriptors

H.3.3 [Information Systems]: Selection process, Retrieval models

## General Terms

Algorithms, Experimentation

## Keywords

Keywords are your own designated keywords.

## 1. INTRODUCTION

Today, millions of news, reports, books, scientific articles are easily accessible in the electronic form almost any place in the world. Streams of information are created by news agencies. As a result, the creation of effective methods for automatic categorization of text documents based on their content is ever more important.

People classify text documents using their knowledge about a certain domain in order to find important facts and meaningful connections between them. In some cases, some external knowledge is required to establish proper relationships among the

facts. In this paper, we propose to use knowledge represented in the form of ontology for categorizing documents. The novelty of this approach is that it does not require a training set of documents, as we categorize text to a given ontology by using the knowledge it contains. The information from the analyzed document is enriched with external ontological knowledge. The categorization process depends solely on entities, named relationships, and the taxonomy of classes represented in the ontology.

“Classical” categorization methods are based on machine learning and probabilistic approaches. A good review of the current categorization methods is presented in [24]. Commonly known methods such as Naïve Bayes [15] or Support Vector Machines [30] have been found to give very good results. Many other methods have been created, and modifications and improvements are constantly introduced. Yet, all of them rely of some kind of a training set of pre-classified documents to perform classification of previously unseen documents into predefined categories.

We decided to relinquish the training set requirement in favor of using only ontological knowledge for categorization; in this way the ontology effectively *becomes* the classifier. The World Wide Web Consortium (W3C) describes an ontology using the following words: “An ontology defines the terms used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information (...) Ontologies include computer-usable definitions of basic concepts in the domain and the relationships among them (...).”<sup>1</sup> We chose to use an ontology created from Wikipedia<sup>2</sup>, as Wikipedia already covers and describes a wide area of general knowledge, and includes computer usable definitions of interconnected and categorized concepts. Moreover, it contains an extensive network of human-understandable categories that can be leveraged in the categorization process.

The proposed approach tries to find a semantic similarity between a document and some fragment of the ontology that describes a certain category. To achieve it, the analyzed document must be transformed into a structure that is similar to the ontology. This process focuses on the creation of a semantic graph from the document, and employs entity matching and relationship identification. The semantic graph is then used to measure the document’s semantic similarity to the categories defined in the ontology.

---

<sup>1</sup> <http://www.w3.org/TR/webont-req/#onto-def>

<sup>2</sup> <http://en.wikipedia.org>

## 2. MOTIVATION

Typically, classes in the ontology are organized into hierarchies. A class defines the types of properties common to the individual objects within the class. Furthermore, the classes are interconnected by relationships, defining some form of semantic interdependence (relationships are also regarded as properties) [26]. Class hierarchies and class relationships form the *schema level* of the ontology. The individuals (instances) and links among them (instances of relationships) form the so called *ground level* of the ontology. RDF/S [5] and OWL [19] are two popular ontology specification languages in use today.

A comprehensive, well populated ontology with classes and relationships closely modeling a specific domain represents a vast compendium of knowledge in the domain. Recently, such ontologies have been used in a variety of semantics-based applications, ranging from business analytics [25] to semantic data integration [6].

We believe that the knowledge represented in a comprehensive ontology can be leveraged to identify topics in a text document, provided the document is thematically related to the ontology domain. Furthermore, the concepts in the, especially if they are organized into hierarchies of higher-level categories, may be used to identify the category that best classifies the document based on its content.

As an example, let us assume that we have a well-defined and comprehensive ontology containing knowledge about a variety of news topics, including business, politics, travel, sports, and others. We will assume that the ontology includes a wide variety of concepts of each topic. For example, in business, the ontology contains many business terms, such as *takeover*, *time to market*, *business alliance*, *takeover*, *outsourcing*, as well as relationships, for example specifying that a *management buyout* is a form of *acquisition* where company's *managers* acquire a large part or all of the company. Furthermore, let us assume that the ontology contains all the relevant instances, such as companies (General Electric, Ford, Nokia), products, brand names, as well as the management personnel of each company, and links among them. For example, such ontology would specify that Alan Mulally is the CEO of Ford and that Intel Corporation's main products are microprocessors and motherboard chipsets. We will also assume that our ontology classes are organized into a hierarchy of higher level classes that group our concepts and instances into a number of broad categories, such as semiconductor companies, technology companies, companies, so on.

Now, consider a news article describing activities of an automotive company. We believe that such an article most likely contains a number of occurrences of concepts and/or individuals represented in our ontology. Such occurrences are known as named entity occurrences and may offer initial clues that the document is, in fact, about automotive business. First, we may be able to identify several named entity occurrences in the text of the document. Consider the following document excerpt:

*Ford Motor Co., is in the process of selling Jaguar and Land Rover, according to Ford CEO Alan Mulally. The Associated Press is reporting the American car manufacturer is in active discussions with potential buyers. Although Mulally has not revealed the names of any potential buyers, India Tata Motors is believed to be in contention.*

*Jaguar may be more difficult to sell than Aston Martin or Land Rover. The newly released Jaguar XF is so critical to that division success, that its failure may sink them. The XF is a modern move for Jaguar, which has hung its hat on its retro heritage for the last four decades. That car goes on sale in March 2008.*

*Ford is also rumored to be looking for someone to buy up the Volvo car division, but the U.S. corporation has denied those statements. Just last week, newspapers in Sweden were calling Volvo AB, the owner of Volvo trucks and original owner of Volvo cars, a potential bidder for the automotive division.*

*In response, Volvo AB CEO Leif Johansson told Bloomberg News. "We are of course very concerned and interested in what happens with Volvo cars, but not such that we would be an investor there."*

We could identify many entities (underlined) and, on the basis of our ontology, we could establish a number of relationships among the identified entities. For example, the facts that Alan Mulally is the CEO of Ford Motor Company, Ford is the owner of Jaguar, Land Rover and Volvo cars divisions, and that Leif Johansson is the CEO of Volvo AB, which owns Volvo trucks. Also, we could discover that the entity (concept) investor is a business term, and that Ford, Volvo, and Land Rover are motor vehicle manufacturing companies.

Based on the above information, we could construct a graph made of the entities in the document and the relationships identified in the ontology. Such a graph, which we call a semantic graph, represents the thematic content of the document, could be used to determine the main topic of the document or establish its category.

Note, that the same phrase (or a single word) may refer to several distinct entities in the ontology. For example, Jaguar, recognized in the above text, may refer to the car manufacturer, the animal of the *Felidae* family, or perhaps British heavy-metal band. As a consequence, it is possible that more than one semantic graph could be created for the entities identified in the document. We can view each of such graphs as representing a plausible interpretation of the thematic content of the document. In case when several semantic graphs can be created, the one which presents the best fit with the ontology would be selected as the dominant semantic graph.

Finally, on the basis of the dominant semantic graph we could establish the overall category of the document, in that we might be able to identify one, or perhaps a small number of categories that classify all, or most of the entities and relationships in the dominant semantic graph.

Another important observation is that in one of the sentences in the above text: "... and that Leif Johansson is the CEO of Volvo AB, which owns Volvo trucks" we could identify not only the (underlined) entities, but we might be able to recognize the direct relationships "is chief executive of" between Volvo AB and Leif Johansson, and "owns" between Volvo AB and Volvo trucks. Natural Language Processing (NLP) techniques could be used to recognize such relationships directly in the document and then used to provide more evidence that the semantic graph fits within our ontology.

We believe that comprehensive ontologies will soon be available for numerous domains. Recently, several ontologies have been created in the area of biology [3], medicine [7], and culture [18]. Work is underway on creating an ontology based on Wikipedia, a popular Web encyclopedia. An RDF version of Wikipedia described in [2] is an interesting intermediate step towards this goal. We believe that the comprehensive knowledge in such a Wikipedia-based ontology can be used as the basis for training-less text categorization.

### 3. RELATED WORK

Text categorization is a well-researched problem in computer science. Perhaps, the existing categorization methods might be improved with the use of external knowledge that is not present, or is virtually impossible to extract from the source document itself. One of the widely used sources of external knowledge for text categorization is WordNet [1]. It is a network of related words, organized into synonym sets, where each of the sets represents one lexical underlying concept. With its help, it is possible to unify similar words into a single phrase or concept used later during categorization. WordNet has been successfully used both in text categorization [22] and clustering [10].

Text categorization using semantic concepts is a step away from simple word or phrase-based categorization towards a semantics-based classification. Latent Semantic Analysis [14] offers an attractive way to transition from the word-space to the concept-space of related phrases. The extracted concepts can be effectively applied to classical categorization, as presented in [11].

Ontologies organize facts and knowledge in a meaningful way. Due to an increasing availability of comprehensive ontologies, their use in information retrieval became an intense topic of research. Similarly to WordNet, ontologies can be used for word sense disambiguation or unification of the vocabulary, as presented in [4]. Ontology knowledge represented by entities inter-connected by named relationships and the defined taxonomy of classes may become a much more powerful tool in information retrieval. Here, categorization can be based on the recognized and disambiguated named entities, as presented in [9] and [27]. Traditional classification methods can also be enriched with the ontology-based information concerning the co-occurring entities [17] or whole neighborhoods of entities [8].

Recently, Wikipedia became an important resource for performing a variety of text analysis tasks. It contains a vast amount of encyclopedic knowledge that is richly interconnected and categorized. Each page can be treated as a named entity, meaningfully connected to other entities and partially categorized. Although in its current form it is not an ontology and does not have a proper taxonomy, initial work has been started on creating Semantic Wiki [29] that would fulfill these requirements. Page categorization in Wikipedia is a mixture of taxonomy and collaborative tagging and can be treated as a thesaurus [31]. The category graph extracted from Wikipedia can be used for the identification of document topics, as described in [23]. The work presented in [20] shows good results in creating a taxonomy from the current Wikipedia categories.

NLP tasks can also benefit from using Wikipedia knowledge. Examples include semantic relatedness [28], based on Wikipedia entries that successfully replace parts of WordNet and co-

reference resolution [21] based on Wikipedia entities and their categories.

### 4. CATEGORIZATION ALGORITHM

Text categorization, also known as text classification, or topic identification, involves the activity of labeling natural language texts (documents) with thematic categories from a predefined set. In supervised text categorization, it is assumed that the categories are just symbolic labels and that no definition of their meaning is available, either procedurally or declaratively [24].

Our text classification algorithm is significantly different from the traditional supervised text classification methods. It does not require classifier training (and a training set of documents), and relies *only* on the domain knowledge represented in the ontology to perform text categorization. Furthermore, each classification category is specified in terms of ontology classes and may be defined as a single class, a subset of the class taxonomy, a list of classes with a specified importance, or perhaps as a combination of taxonomy and list of classes.

Our categorization algorithm consists of three main steps: the construction of the semantic graph, the selection and analysis of the thematic graph, and the categorization of the selected thematic graph. The outline of the algorithm is presented below:

#### *Semantic graph construction*

- Named entities identification – phrases describing entities (entity labels) in the ontology are matched in the text; for each located phrase, all associated entities are added as nodes to the created semantic graph (in what follows, we will treat nodes in the graph and entities as synonymous); each node is assigned an initial weight based on the strength of the match.
- Connectivity inducing – edges between the nodes in the semantic graph are created based on the relationships existing in the ontology and connecting the entities corresponding to the nodes; each edge is assigned a weight based on the importance of the relationship in the ontology schema.
- Information propagation – node weights are propagated to their neighbors in order to establish the most authoritative entities in the graph.

#### *Thematic graph selection and analysis*

- Connected component identification – connected components in the semantic graph are identified, treating the graph as undirected.
- Dominant thematic graph identification – the largest and most important connected component of the semantic graph is selected as the dominant thematic graph.
- Core selection – the most important and/or central entities in the thematic graph are identified; they form the core of the thematic graph.

#### *Dominant thematic graph categorization*

- Class assignment – each entity in the dominant thematic graph is assigned a set of classes, according to the entity's classification in the ontology (we assume that an entity may belong to multiple classes); for each class in the set, its depth in the ontology class hierarchy is recorded.

- Ontological classification – starting with the classes assigned to the authoritative and/or central entities (we say, *covering* the authoritative and/or central entities), ascend in the ontological class hierarchy until a set of parent classes is located that covers a significant portion of entities in the dominant thematic graph. Each class in the new set of (higher level) classes is ranked according to (i) the weight of the entities it covers, (ii) the percentage of the covered entities in the dominant thematic graph, and (iii) the distance in the class hierarchy to the covered entities.
- Categorization – the target categories are defined as ontology class (sub-)hierarchies, or just lists of classes; each class in the set located in the previous step is determined to belong to one or more categories (as a member of the hierarchy or the list); the weight of the classes is used to determine the best category for the document.

The proposed training-less text categorization method is based on the semantics offered by the ontology. It should be noted that meaningful results can only be achieved under the following assumptions (1) the entities in the ontology have proper label(s) that can be used for entity spotting in categorized documents, (2) the entity labels are in the same language as the categorized documents, (3) the used ontology has a rich instance base of interconnected entities, (4) the entities are classified according to a taxonomy included in the ontology, (5) the analyzed text is related to one (or more) knowledge domains represented in the ontology.

#### 4.1 Construction of the semantic graph

Before a text document can be categorized using this method, it must be converted into a graph structure. The conversion process superimposes a subset of the ontology (entities and relationships among them) on the analyzed document.

Each entity in the ontology that can be recognized in the document has one or more descriptive literals associated with it. These are values of certain key properties that can be used to name or identify the entity. Examples of such properties include the entity’s name (known as its label) or its alternate names (usually assigned as aliases). We assign different importance weight to each property type connecting an entity and its descriptions. In general, the names are more important than aliases or shortened descriptions. *Entity spot phrases* are created from values of the selected identification properties; we do not assume that names or aliases uniquely identify entities. Stemming and stop words removal may be applied to both the document text and spot phrases.

A spot phrase can be matched in several places in the document. This indicates that the same entity is mentioned in different places of the document and therefore may be more important for the document topic. Multiple occurrences of the same phrase describing an entity cannot be treated in the exact way as term frequency in traditional classification methods, as it is likely that a frequently appearing entity is connected to the document content stronger than other, less frequently occurring entities. However, the weight (importance) of an entity should not depend linearly on its frequency. We use the following formula for establishing the initial entity weight:

$$w = 1 - \frac{1}{1 + \sum_{i=1..n} p_i * s(l_i, sp_i)}$$

In the formula,  $w$  is the initial entity weight and  $n$  represents the number of matches of the spot phrases associated with the entity. The term  $p_i$  is the weight of the property connecting the literal  $l_i$ , which is associated with spot phrase  $sp_i$ , to the entity  $w$  in the  $i$ -th match. The function  $s$  measures the similarity between the spotted phrase  $sp_i$  and the original literal  $l_i$  in the ontology. If word stemming and stop word removal is not used, the similarity degenerates to simple string equality. For similarity using stop word removal we accommodated following formula:

$$s(l_i, sp_i) = \frac{nw(sp_i)}{nw(l_i - \{ "a", "an", "the" \})}$$

Similarity  $s$  is a ratio of the number of words in the spot phrase  $sp_i$  (created by removing stop words from literal  $l_i$ ) to the number of words in the original literal  $l_i$ , not counting the articles “a”, “an” and “the”. Note that the stop list is much longer and also includes the above articles.

Note, that a matched spot phrase can be associated with multiple entities in the ontology. This may lead to an incorrect identification of many entities (false positives) and the number of recognized entities may even be higher than the number of phrases spotted in the document. However, there is no disambiguation performed at this stage and all matched entities are included as nodes in the constructed semantic graph. Incorrectly matched entities will be removed at a later stage.

The conversion of the analyzed text into a semantic graph requires the addition of the relationships between entities. In the proposed algorithm we add relationships only existing in the ontology. If a pair of entities recognized in the document is connected by a relationship in the ontology, an edge, corresponding to the same relationship, is added to the created semantic graph. This process can be seen as adding domain knowledge from the ontology to the analyzed document. The *induced relationships* add semantic connections and domain-specific interpretation to the recognized entities, providing an explanation of their co-occurrence in the document, and the resulting graph is called a *semantic graph*.

Our algorithm utilizes the created semantic graph to identify the most important entities, even if they occur infrequently in the document. To facilitate the search for such important entities, we assign varying importance weights to different named relationships. One of the possible automatic weight assignments is based on relationship rarity in ontology. The “Hubs and authorities” algorithm [13] is later used to propagate the weights and locate the authoritative entities.

#### 4.2 Thematic graph selection and analysis

A document converted to a semantic graph may contain many entities added during the matching phase that are unrelated, weakly related to the document topic, or even incorrectly matched in the document context. Additionally, it is possible that different sections of the document concern different topics.

This step attempts to isolate a sub-graph of the semantic graph that are related to the main topic of the document. It is based on the ontology semantics that assumes that entities related to a

single topic are (closely) connected in the ontology, and unrelated entities are placed far-apart or even not connected. The semantic graph created from the document represents only fragments of the original ontology, and may be composed of several connected components. We iterate over the connected components and, following the semantic closeness assumption, try to find the dominant one that best describes the main topic of the document. The component with the largest number of nodes and the highest total of the entity weights becomes the *dominant thematic graph* for the document.

The selected dominant thematic graph effectively defines the thematic interpretation of the document which will be used in further analysis. It also removes any entities and parts of the graph that are not related to the selected topic. In cases when multiple entities were matched by a single phrase, fixing the interpretation effectively disambiguates them, and leaving only the relevant ones.

Selection of the core entities is the last step in the thematic graph analysis. Naturally, the authoritative entities fall into this category. Additionally, we compute the centrality score to find the most central entities in the structure of the dominant thematic graph. We use the geodesic closeness formula:

$$Centrality(v_i) = \frac{1}{\sum_j d(v_i, v_j)}$$

where  $d(v_i, v_j)$  is the shortest path distance between vertices  $v_i$  and  $v_j$  in the undirected graph. The most central nodes become the landmarks of the analyzed topic. Note, that they may be different than the authoritative nodes.

The core of the dominant thematic graph includes the top authoritative nodes and the most central ones. In our experiments, we limited their number to 10% of the nodes in the dominant thematic graph. To ensure the presence of both types of nodes and to have a set minimum core size, we require that at least 3 entities from each type be present.

### 4.3 Thematic graph categorization

The dominant thematic graph serves as the source of information in the categorization process that uses the class taxonomy from the ontology schema. For each entity in the thematic graph, we find all classes assigned to the entity in the ontology schema, and all super-classes, using the transitivity of the sub-class relationship. In this process, a set of reachable classes is created, where each of found class has a list of assigned entities together with hierarchical distance to them. The distance between the entity and its immediate class (as defined in the ontology) is 1.

The categorization process focuses on finding the best class describing the thematic graph and shifts the attention from the instances (entities) to the ontology class taxonomy. In ranking the classes, we consider which one (1) covers (is a super-class of the entity) the largest number of important entities in the thematic graph, (2) is very close (in hierarchical distance) to the entities, (3) maximizes the coverage of the core entities and other entities in the thematic graph.

The selected class should offer the best fit for whole thematic graph (or its large part). Taking into consideration the properties

of the best coverage class, we use the following formula for calculating the class score:

$$s_{C_i}(h_{max}) = 1 - \left(1 - \frac{1}{1 + \sum_j \frac{w_j}{h(C_i, e_j)^2} + \sum_k \frac{w_k}{h(C_i, e_{Ck})}}\right)$$

where  $s_{C_i}$  is the categorization score for class  $C_i$ , that includes reachable entities  $e$  up to depth  $h_{max}$ ;  $e_j$  and  $e_{Ck}$  represent respectively entity and core entity reachable up to depth  $h_{max}$  from class  $C_i$ ;  $w_j$  and  $w_{Ck}$  are weights of entity  $e_j$  and core entity  $e_{Ck}$ ;  $h(C_i, e)$  is the hierarchical distance between category  $C_i$  and the covered entity  $e$ . The first summation is over all of the entities reachable from category  $C_i$  up to depth  $h_{max}$ , while second include only core entities.

The classes, ranked according to their scores, represent the categorization of the document, relatively to the taxonomy in the ontology schema. The final step of the document categorization requires matching between the internal categorization and user-defined topic categories. There are many possible approaches to match the assigned taxonomy classes with selected topics. We suggest following different approaches that depend on the method used in defining the topic categories.

The proposed score computation favors the classes that are closest to the thematic graph. In case the final topics are defined as parts of the taxonomy, we should favor matching to the lowest category in the hierarchy as the most specific one. Final categorization score  $s_{C_i}$  for class  $C_i$  should be increased as the class depth increases. It can be done linearly, exponentially or in other method that incorporates user's interest in matching deep categories.

In case the class hierarchy is similar to that found in Wikipedia, where the defined categories instead of a clear taxonomy form a thesaurus, such an approach will likely give unsatisfactory results. In a thesaurus, the further we move down from the original category, the less relevant the matched category becomes. The categorization score should be modified opposite to the taxonomical case.

Assigning external category is based on matching modified highest ranked classes to category definition. Starting from the class with the highest score, assign it to all appropriate categorization topics and increase their weight accordingly. If only one topic has the highest score, select it as the categorization result. Otherwise, take next class in the ranking and repeat the process until only one topic has the highest total score, or there are no more ranked classes to select.

## 5. EXPERIMENT SETUP

Our experiments focused on testing the proposed training-less ontology-based categorization method and comparing its accuracy to a base-line traditional method. We have chosen the well known BOW toolkit [16] as the representative implementation of one of the traditional categorization methods. As in other supervised text categorization methods, it requires a document training set.

The ontology used in our classification method was created from the English version of Wikipedia, using a similar approach as the one presented in DBpedia [2]. Our text corpora consisted of news articles downloaded from the RSS feeds from CNN.com.

Our prototype implementation of the ontology-based categorization uses BRAHMS [12] as the backend RDF storage to ensure highly efficient handling and querying of the large RDF ontology created from Wikipedia.

The details about the experiment setup and the used methodology are presented in the subsequent sections.

## 5.1 Wikipedia ontology

We used the English Wikipedia dump from 2007-09-08 as the source for creating the ontology. We utilized a modified DBpedia approach to create an RDF/S ontology from Wikipedia. Our modifications related to handling of the extraction of the templates included in a typical Wikipedia entry and assignment of literal values.

DBpedia focuses on special handling of Infoboxes. Entities mentioned in an infobox are linked to the source entity by named relationships derived from the Infobox description. There is no special handling of general templates included in the page. Such templates become separate entities linked from the source entity, which further link to other pages. On a Wikipedia page, such templates are included in the page source and entities they point to are directly linked from the page. For example, in the article about “Atlanta Hawks,” we have the “current roster” section, which is internally represented as a template. In the created ontology we shortcut such links in the included templates without creating an intermediate entity.

The literal values, such as entity names, redirections, and disambiguation are very important for creating phrases to spot entities in the document. We have created separate named relationships to distinguish among the direct names of entities (page names in Wikipedia), redirections (redirection pages), and entities names included in the disambiguation pages. Wikipedia also utilizes a convention of disambiguating entity names by adding contextual information enclosed within parentheses and listed after the entity name, e.g., “Jaguar”, “Jaguar (car)” and “Jaguar (band)”. Such full phrases do not exist in documents, as the context of the document provides enough information for human to properly disambiguate the entity. For entities with such names, we create a shorter literal by omitting the context information and add it as an alternate, shorter name, using specific property to distinguish it from the full name.

Our derived instance base of the RDF/S Wikipedia ontology contains 2,062,198 instance entities that contribute to a highly connected graph of 67,279,865 statements, and 4,409,200 literals

that can be used for entity matching. On average, each entity has assigned 2.85 literals using different relationships. The schema part has 311,908 classes (Wikipedia categories) organized in 532,191 statements, mostly describing semi-hierarchical dependencies among the classes. Each entity, on average, has been assigned to 2.64 classes.

## 5.2 Text corpora

We tested the proposed categorization method on the CNN news articles (www.cnn.com) obtained from CNN RSS feeds between 2007-07-03 and 2007-09-04. The prepared corpora consist of 2,590 news articles assigned to 12 categories, each associated with a single RSS feed. For comparison with a traditional, probabilistic categorization method, we divided it into a 50/50 split, where the training and testing sets had 1,295 documents each. The selected categories with the split details are presented in Table 1.

The choice of selecting a set of recent news articles as the text corpora for the categorization tests relates to choosing Wikipedia as categorization ontology. Since Wikipedia contains up-to-date general knowledge, the recent news articles should be a suitable match for the initial tests. Furthermore, the CNN news categories can be relatively easily matched to the general Wikipedia categories.

## 5.3 Category mapping

The proposed categorization method ranks the document against the internal ontology classes. To classify a document into a given ontology, no additional information or mappings are necessary. However, classification of the document into *external* categories, such as the ones used by CNN, requires that each category be defined in terms of the internal ontology classes.

In this experiment, we used the ontology derived from Wikipedia, where the category pages were converted into the internal ontology classes. Although organized as a hierarchy, they did not form a taxonomy, since the Wikipedia categories are not based on the strictly defined relationship. In fact, the Wikipedia category pages form a *thesaurus* of concepts, and we have leveraged this feature in creating a mapping from the CNN categories to Wikipedia classes.

For each CNN category, we have manually selected the main concepts from among the Wikipedia categories (roots) and added their subcategories up to the depth of 3. To minimize the ambiguity, we did not allow the selected root categories to be assigned into more than one category mapping. This have added a

Table 1. CNN corpora split with Wikipedia assigned root categories.

CNN category	Training	Testing	Wikipedia root classes	Number of
Education	4	7	Category:Education	1467
Health	91	87	Category:Health	1505
money_autos	37	26	Category:Automobiles	1350
money_companies	271	275	Category:Business, Category:Economics,	2509
money_taxes	15	12	Category:Accountancy, Category:Taxation	146
politics	171	167	Category:Politics, Category:Politicians	7746
science_and_space	35	27	Category:Science, Category:Space	833
sport_mlb	143	171	Category:Baseball, Category:Major League Baseball	1710
sport_nba	139	122	Category:Basketball,	2438
sport_nfl	203	222	Category:Football, Category:National Football League	8251
sport_nhl	93	100	Category:Hockey, Category:National Hockey League	1858
travel	93	79	Category:Travel	714

level of protection against adding the same categories (and their subcategories) into multiple mappings. The depth limit was set due to the fact that we were dealing with a thesaurus, and the deeper we moved from the original concept, the sub-concepts became less and less connected by the “subclass of” relationship. The Wikipedia categories selected for roots of the CNN mappings are presented in Table 1.

## 5.4 Reference categorization method

We have selected the Naïve Bayes classification method available as part of the BOW toolkit to be the baseline used for comparing the categorization accuracy. At first, we used the classical approach, where the training set came from the same corpora as the classified documents. For this purpose, we trained BOW on the training part of our CNN split.

In another experiment, we utilized the documents from Wikipedia in the Naïve Bayes categorization. We prepared a training set of documents selected from Wikipedia for each of the mapped CNN categories. Here, the documents (Wikipedia entries) were selected for the training set of a given CNN category, if their Wikipedia categorization belonged to the selected mapping. The Wikipedia categories selected for all CNN category mappings include over 400,000 documents (pages). To prevent overtraining, we limited the size of the training set for each category to 2,000 documents, chosen randomly from the available set. To check the consistency of the categorization results, 10 different training sets were created and tested. We believe that this experiment offered a better comparison with the direct ontology-based classification, as both the traditional (probabilistic) and ontology-based classifiers used the same source of information for the categorization task.

## 6. EXPERIMENT RESULTS

We performed the following five experiments:

1. Naïve Bayes categorization of our CNN document split using the training set from the same source as the categorized documents (Table 2),
2. Naïve Bayes categorization of our CNN documents using the training set derived from Wikipedia articles assigned to the classes in the mapped categories (Table 3),
3. Naïve Bayes categorization of a set of Wikipedia documents derived from the category mapping, using a different set of Wikipedia documents for training, (Table 4)
4. ontology-based categorization of our CNN documents to the mapped categories and no training set was used (Table 5),
5. ontology-based categorization of a set of Wikipedia documents derived from the category mapping and no training set was used (Table 6).

The first experiment has been used as the baseline in how well a classical classification method can perform on given categories and text corpora. The second and third experiments compare the traditional and ontology-based methods, when the classifiers use the documents from the same source, outside of a given test documents. The last two experiments serve as a verification of the created category mappings and provide some insight into self-classification of Wikipedia articles to their original categories.

The best results were produced by Naïve Bayes categorization trained on CNN split documents. With 1,295 training and 1,295

testing documents from the same source, the overall accuracy was over 92%.

Categorization of the CNN corpora by Naïve Bayes categorization using a subset of Wikipedia documents as the training set achieved only 73% accuracy. The testing set, as in the previous experiment, consisted of 1,295 documents, but the training had assigned 22,154 Wikipedia documents. Our ontology-based method that used our prepared CNN category mappings reached a similar accuracy of 80%. These results may be regarded as more comparable, as the knowledge (documents) from the same source was used for the categorization.

The last two tests were performed for testing Wikipedia self classification, according to our mapping of categories. The Wikipedia pages selected for both the testing and training sets for the Naïve Bayes categorization. We have found that among 10 different trials, the classification was steadily achieving around 83% accuracy. Our ontology-based categorization of Wikipedia articles had only 67% accuracy.

## 6.1 Results analysis

The proposed ontology-based categorization method achieved overall good results on the tested CNN corpora. The accuracy was comparable to Naïve Bayes when trained on the documents from the ontology. Still, Naïve Bayes outperformed our method when it was trained on the articles from the original corpora.

In the experiments involving the categorization of Wikipedia pages to predefined CNN categories, the Naïve Bayes method remained at the same accuracy level, whereas the ontology-based categorization results dropped somewhat.

We investigated the potential sources of misclassification problems in both the CNN and Wikipedia corpora. Analysis of several articles and created thematic graphs, together with the ranked categories revealed following causes of misclassifications:

- the difference between the article’s thematic content and the assigned category,
- the graphs created from the longer Wikipedia articles have been multi-thematic and the select dominant thematic graph was outside of the assigned category, the created mapping between the CNN and Wikipedia categories were too broad and imprecise,
- an unevenly developed structure of Wikipedia link and category for different domains.

The fact that the training set categorization has been based on the reader’s *perceived* interest is the most common cause for the misclassifications. The ontology-based categorization analyzes the document content in order to create a thematic graph, and then finds its best fit into the ontological knowledge. On the other hand, the categories assigned by CNN mainly reflect the reader’s perceived interest. Our Wikipedia-based ontology contains encyclopedic knowledge, which describes basic facts and their relationships. It does not favor any specific types of entities such as people, companies, or places. The (human assessed) perceived interest in the article may decide the article’s category solely on a single type of high-interest entity, and not on the thematic content of the document.

As an example, consider an article about cardio-vascular health problems of a certain politician. From a reader’s perspective, the article belongs to politics, as the politician is the main point of

interest. On the other hand, the majority of the document content is about the disease, treatment, or perhaps recovery. In the analysis of the created semantic graph, the politician most probably will not become one of the core entities, and the graph core will concentrate on medical issues. This will result in the final categorization into the health domain.

The imprecision in the category mapping for the categorization of CNN news articles was another source of misclassifications. We tried to utilize a simple approach based on selecting root

Wikipedia categories and including additional related categories. In most cases, the categories included in the mapping were strongly related to the root category. As the number of categories grew exponentially with the distance from the root, it was not uncommon that additionally included categories had little in common with the root one. For example, in the politics category with the root Politics, we can find categories including Locksmithing, Olympics, or Hydrology. It is possible to manually refine the categories and achieve better results, but it goes against

**Table 2. Naïve Bayes categorization (BOW implementation) results of CNN document split with CNN training set.**

Correct: 1220 out of 1295 (94.21 percent accuracy); Confusion details, row is actual, column is predicted

Classname	0	1	2	3	4	5	6	7	8	9	10	11	total	correct
0Education	2	.	.	.	.	2	.	.	.	.	.	.	4	50.00%
1Health	.	80	.	4	.	2	.	.	.	2	.	3	91	87.91%
2money_autos	.	.	16	20	.	.	.	.	.	.	.	1	37	43.24%
3money_companies	.	1	4	263	.	2	.	.	.	.	.	1	271	97.05%
4money_taxes	.	.	.	9	4	2	.	.	.	.	.	.	15	26.67%
5Politics	.	2	.	.	.	169	.	.	.	.	.	.	171	98.83%
6science_and_space	.	.	.	.	.	.	33	.	1	.	.	1	35	94.29%
7sport_mlb	.	.	.	1	.	.	.	140	.	2	.	.	143	97.90%
8sport_nba	.	.	.	.	.	.	.	1	135	3	.	.	139	97.12%
9sport_nfl	.	.	.	.	.	1	.	.	.	202	.	.	203	99.51%
10sport_nhl	.	.	.	1	.	.	.	.	2	.	90	.	93	96.77%
11Travel	.	1	.	2	.	4	.	.	.	.	.	86	93	92.47%

Percent\_Accuracy average 94.21 stderr 0.00

**Table 3. Naïve Bayes categorization (BOW implementation) results of CNN document split with Wikipedia documents training set.**

Correct: 1949 out of 1295 (73.28 percent accuracy); Confusion details, row is actual, column is predicted

Classname	0	1	2	3	4	5	6	7	8	9	10	11	total	Correct
0Education	7	.	.	.	.	.	.	.	.	.	.	.	7	100.00%
1Health	23	55	.	1	1	.	3	.	.	.	.	4	87	63.22%
2money_autos	.	.	23	3	.	.	.	.	.	.	.	.	26	88.46%
3money_companies	1	11	16	169	74	.	.	.	.	.	.	4	275	61.45%
4money_taxes	.	.	.	.	12	.	.	.	.	.	.	.	12	100.00%
5Politics	11	4	.	2	40	100	.	.	.	.	.	10	167	59.88%
6science_and_space	1	.	.	.	.	.	22	.	.	.	.	4	27	81.48%
7sport_mlb	1	1	.	3	5	.	.	155	.	.	.	6	171	90.64%
8sport_nba	2	.	.	3	4	.	.	.	99	1	.	13	122	81.15%
9sport_nfl	13	1	.	9	7	2	.	.	8	160	.	22	222	72.07%
10sport_nhl	5	.	1	7	5	.	.	.	2	.	75	5	100	75.00%
11Travel	1	1	.	2	3	.	.	.	.	.	.	72	79	91.14%

Percent\_Accuracy average 73.28 stderr 0.00

**Table 4. Naïve Bayes categorization (BOW implementation) results of Wikipedia documents using Wikipedia pages as training set.**

Correct: 18451 out of 22154 (83.29 percent accuracy); Confusion details, row is actual, column is predicted

Classname	0	1	2	3	4	5	6	7	8	9	10	11	total	Correct	
0education	1655	12	1	24	42	60	109	.	1	1	.	.	47	1952	84.78%
1health	232	1243	6	42	26	30	198	.	1	4	3	125	1910	65.08%	
2money_autos	11	3	1665	63	5	13	37	.	.	5	1	132	1935	86.05%	
3money_companies	73	12	36	1303	147	33	42	.	1	3	3	220	1873	69.57%	
4money_taxes	4	.	.	4	869	31	4	1	.	3	1	11	928	93.64%	
5politics	147	3	2	39	104	1425	12	1	1	14	4	116	1868	76.28%	
6science_and_space	289	131	24	38	58	24	1218	.	.	.	.	1	147	1930	63.11%
7sport_mlb	2	.	.	3	2	.	.	1933	27	29	8	9	2013	96.03%	
8sport_nba	5	.	.	26	.	6	.	13	1883	37	14	14	1998	94.24%	
9sport_nfl	4	.	1	11	.	9	.	5	18	1885	12	9	1954	96.47%	
10sport_nhl	8	.	1	18	.	1	.	2	49	39	1742	43	1903	91.54%	
11travel	48	24	21	71	34	17	39	1	2	2	1	1630	1890	86.24%	

Percent\_Accuracy average 83.29 stderr 0.00

**Table 5. Ontology-based categorization of CNN document split using Wikipedia ontology with prepared category mapping.**

CATEGORY	0	1	2	3	4	5	6	7	8	9	10	11	12	total	correct
0Education	6	.	.	.	.	1	.	.	.	.	.	.	.	7	85.71%
1Health	2	70	.	3	.	4	4	.	.	.	.	4	.	87	80.46%
2money_autos	.	.	17	8	.	.	1	.	.	.	.	.	.	26	65.38%
3money_companies	.	20	10	213	19	2	5	1	.	.	.	5	.	275	77.45%
4money_taxes	.	.	.	3	8	1	.	.	.	.	.	.	.	12	66.67%
5Politics	.	6	.	8	.	148	3	.	.	.	.	2	.	167	88.62%
6science_and_space	1	1	.	1	.	1	21	.	.	.	.	2	.	27	77.78%
7sport_mlb	.	2	.	6	.	3	1	153	.	.	.	6	.	171	89.47%
8sport_nba	.	3	.	7	.	12	3	.	91	1	.	4	1	122	74.59%
9sport_nfl	.	3	1	7	.	16	4	.	.	185	.	6	.	222	83.33%
10sport_nh	.	.	1	7	.	3	2	.	1	3	77	6	.	100	77.00%
11Travel	.	5	.	7	.	9	10	.	.	.	.	48	.	79	60.76%
12Unknown	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0.00%
Classified documents : 1295															
Correctly classified : 1037															
Achieved accuracy : 80.077															

**Table 6. Ontology-based categorization of Wikipedia documents using Wikipedia ontology with prepared category mapping.**

CATEGORY	0	1	2	3	4	5	6	7	8	9	10	11	12	total	correct
0education	922	173	4	191	1	124	449	1	1	4	2	53	27	1952	47.23%
1health	60	1223	3	81	1	81	409	.	1	2	1	34	14	1910	64.03%
2money_autos	18	15	1519	181	.	62	103	.	.	6	1	22	8	1935	78.50%
3money_companies	34	89	42	1187	15	132	165	2	.	2	.	135	70	1873	63.37%
4money_taxes	17	35	1	321	362	132	47	.	.	1	.	4	8	928	39.01%
5politics	105	88	2	152	4	1373	84	2	1	8	2	18	29	1868	73.50%
6science_and_space	60	255	13	190	.	84	1290	1	.	1	.	20	16	1930	66.84%
7sport_mlb	21	3	.	7	.	14	6	1940	3	6	1	2	10	2013	96.37%
8sport_nba	217	7	.	60	.	47	11	24	1536	58	14	12	12	1998	76.88%
9sport_nfl	38	6	2	44	.	109	56	6	3	1659	3	8	20	1954	84.90%
10sport_nh	38	15	2	46	.	102	36	6	16	36	1562	22	22	1903	82.08%
11travel	119	189	26	434	3	235	479	2	2	3	.	333	65	1890	17.62%
12unknown	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.00%
Classified documents : 22154															
Correctly classified : 14906															
Achieved accuracy : 67.284															

our assumed simplicity in preparing the mappings, and this could be potentially viewed as a hidden way of introducing a training set.

Finally, some misclassifications were related to the ontology and Wikipedia itself. Some parts of Wikipedia are much better covered and interconnected than others. Consequently, entities from the better covered regions have a higher chance to be recognized and, due to their high connectivity, create a better thematic graph.

Focusing only on the document content, represented by entities and relationships, can be perceived both as the strength and the weakness of our categorization method. The strength comes from utilizing the background knowledge from the ontology that may not be present in the document. The weakness lies in the very difference between facts and perceived interests, which may require a much more sophisticated mapping or a modification of our algorithm to favor certain types of entities, relationships, or structures. We believe that it may be overcome by using certain NLP methods in building the thematic graph, and providing a more specific and defined context of interest in the classification step of the algorithm.

## 7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a novel approach to text categorization. Our method relies only on the ontological knowledge stored in the domain ontology. The novelty of this method is that it does not require a document training set. As rich and comprehensive domain ontologies become available, the approach proposed in this paper may be a suitable alternative to the traditional text categorization methods. The tests performed on an RDF ontology derived from English Wikipedia demonstrated the practical value and effectiveness of the ontology-based categorization.

However, the categorization based on factual knowledge from the ontology does not always match user's perceived interests. In the near future, we plan to concentrate on defining a categorization context to fill this gap. It will enable specification of more complex categories than presented here simple category mapping.

Another direction of future work lies in enriching the thematic graph with more semantic information extracted from the analyzed document and utilizing this information in the categorization process. We plan to investigate how NLP methods, combined with the ontological knowledge can help in discovering

named relationships between entities in the document. Such information would significantly influence certain interpretation contexts by either strengthening already known relationships between entities, or by supplying additional knowledge not yet existing in the ontology.

## 8. REFERENCES

- [1] *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- [2] Auer, S. and Lehmann, J., What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. in *European Semantic Web Conference (ESWC'07)*, (Innsbruck, Austria, 2007), Springer, 503-517.
- [3] Bada, M., Stevens, R., Goble, C., Gil, Y., Ashburner, M., Blake, J.A., Cherry, J.M., Harris, M. and Lewis, S. A Short Study on the Success of the Gene Ontology. *Journal of Web Semantics*, 1 (2).
- [4] Bloehdorn, S. and Hotho, A., Text Classification by Boosting Weak Learners based on Terms and Concepts. in *4th IEEE International Conference on Data Mining (ICDM'04)* (2004).
- [5] Brickley, D. and Guha, R.V. RDF Vocabulary Description Language 1.0: RDF Schema. McBride, B. ed. <http://www.w3.org/TR/rdf-schema/>, 10 Feb 2004.
- [6] Buccella, A., Cechich, A. and Brisaboa, N.R. Ontology-Based Data Integration. in Rivero, L.C., Doorn, J.H. and Ferragline, V.E. eds. *Encyclopedia of Database Technologies and Applications*, Information Science Reference, 2005.
- [7] Eccher, C., Purin, B., Pisanelli, D.M., Battaglia, M., Apolloni, I. and Forti, S. Ontologies supporting continuity of care: The case of heart failure. *Computers in Biology and Medicine* (2006), Jul-Aug; 36 (7-8). 789-801.
- [8] Gabrilovich, E. and Markovitch, S., Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. in *21th National Conference on Artificial Intelligence*, (Boston, MA, USA, 2006).
- [9] Hammond, B., Sheth, A.P. and Kochut, K.J. Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. *Real World Semantic Web Applications*, IOS Press, 2002.
- [10] Hotho, A., Staab, S. and Stumme, G. WordNet improves Text Document Clustering *SIGIR Semantic Web Workshop*, Toronto, Canada, 2003.
- [11] Huang, Y. Support Vector Machines for Text Categorization based on Latent Semantic Indexing *Technical report, Electrical and Computer Engineering Department, The Johns Hopkins University*, 2003.
- [12] Janik, M. and Kochut, K.J., BRAHMS: A WorkBench RDF Store And High Performance Memory System for Semantic Association Discovery. in *Fourth International Semantic Web Conference (ISWC 2005)*, (Galway, Ireland, 2005).
- [13] Kleinberg, J.M., Authoritative Sources in a Hyperlinked Environment. in *ACM-SIAM Symposium on Discrete Algorithms*, (1998).
- [14] Landauer, T.K., Foltz, P.W. and Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes* (25). 259-284.
- [15] Lewis, D.D., Naive (Bayes) at forty: The independence assumption in information retrieval. in *10th European Conference on Machine Learning*, (Chemnitz, DE, 1998).
- [16] McCallum, A.K. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/~mccallum/bow>, 1996.
- [17] Nagarajan, M., Sheth, A.P., Aguilera, M., Keeton, K., Merchant, A. and Uysal, M. Altering Document Term Vectors for Classification - Ontologies as Expectations of Co-occurrence *LSDIS Technical Report*, November, 2006.
- [18] Ossenbruggen, J.v., et. al. Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques. in *Museums and the Web 2007*, (San Francisco, CA, 2007).
- [19] Patel-Schneider, P.F., Hayes, P. and Horrocks, I. OWL Web Ontology Language Semantics and Abstract Syntax <http://www.w3.org/TR/owl-semantics/>, 10 Feb 2004.
- [20] Ponzetto, S.P. and Strube, M., Deriving a Large Scale Taxonomy from Wikipedia. in *Twenty-Second Conference on Artificial Intelligence (AAAI'07)*, (Vancouver, Canada, 2007).
- [21] Ponzetto, S.P. and Strube, M., Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. in *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, (New York, New York, USA, 2006).
- [22] Rosso, P., Ferretti, E., Jiménez, D. and Vidal, V., Text Categorization and Information Retrieval Using WordNet Senses. in *2nd Global WordNet Int. Conf., GWN-2004*, (Brno, Czech Republic, 2004).
- [23] Schonhofen, P., Identifying document topics using the Wikipedia category network. in *ACM International Conference on Web Intelligence*, (Hong Kong, 2006).
- [24] Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34 (1).1-47.
- [25] Semagix. Anti-Money Laundering - CIRAS. [http://www.semagix.com/solutions\\_ciras.html](http://www.semagix.com/solutions_ciras.html).
- [26] Sheth, A.P., Arpinar, I.B. and Kashyap, V. Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships. in Nikravesh, M., Azvin, B., Yager, R. and Zadeh, L. eds. *Enhancing the Power of the Internet: Studies in Fuzziness and Soft Computing*, Springer Verlag, 2003.
- [27] Sheth, A.P., Bertram, C., Avant, D., Hammond, B., Kochut, K.J. and Warke, Y. Semantic Content Management for Enterprises and the Web. *IEEE Internet Computing*, July/August 2002.
- [28] Strube, M. and Ponzetto, S.P., Wikirelate! Computing Semantic Relatedness Using Wikipedia. in *Twenty-First Conference on Artificial Intelligence*, (Boston, MA, 2006).
- [29] Tazzoli, R., Castagana, P. and Campanini, S.E., Towards a semantic WikiWikiWeb. in *Poster Session, 3rd International Semantic Web Conference*, (Hiroshima, Japan, 2004).
- [30] Vapnik, V. *The nature of statistical learning theory*. Springer Verlag, 1995.
- [31] Voss, J. Collaborative thesaurus tagging the Wikipedia way. *ArXiv Computer Science e-prints*, cs/0604036.