Excavating the mother lode of human-generated text: A systematic review of research that uses the Wikipedia corpus

Mohamad Mehdi[a,*], Chitu Okoli[b], Mostafa Mesgari[c], Finn Årup Nielsen[d], Arto Lanamäki[1]

[a] *Computer Science, Concordia University, Montreal, Canada*
[b] *John Molson School of Business, Concordia University, Montreal, Canada*
[c] *Love School of Business, Elon University, Elon, NC, USA*
[d] *DTU Compute, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark*
[e] *Department of Information Processing Science, University of Oulu, Oulu, Finland*

Abstract

Although primarily an encyclopedia, Wikipedia's expansive content provides a knowledge base that has been continuously exploited by researchers in a wide variety of domains. This article systematically reviews the scholarly studies that have used Wikipedia as a data source, and investigates the means by which Wikipedia has been employed in three main computer science research areas: information retrieval, natural language processing, and ontology building. We report and discuss the research trends of the identified and examined studies. We further identify and classify a list of tools that can be used to extract data from Wikipedia, and compile a list of currently available data sets extracted from Wikipedia.

Keywords: information retrieval, information extraction, natural language processing, ontologies, Wikipedia, literature review

1. Introduction

Wikipedia, the largest multilingual wiki-based free-content encyclopedia, is home to more than 32 million wiki pages and 20 million users. Its driving force is the gathering of knowledge by voluntary contributions that cover a wide range of topics. Hundreds of scholarly studies have demonstrated its remarkable competence as a multi-purpose knowledge base [93, 80, 94]. Among these studies, some have focused not on Wikipedia as a phenomenon in its own right, but have rather taken advantage of its enormous collection of semantically-rich, human-generated text and multimedia content to conduct studies where such corpora are needed. Medelyan et al. [78] specifically synthesized the literature on Wikipedia as a textual corpus. However, since their foundational review, there has been extensive research conducted that employs Wikipedia as a data source, and this continues to be an important body of research. Moreover, although they identified many

*Corresponding author
Email address: mo_mehdi@encs.concordia.ca (Mohamad Mehdi)

textual corpus related works, they did not systematically categorize many important research details of these studies, such as the research methodologies employed or the type of the analyzed Wikipedia pages.

This article comprehensively reviews the scholarly literature on using Wikipedia as a textual corpus. It goes beyond the work of Medelyan et al. not only in providing summaries of more recent research, but in carefully analyzing and tracing research trends and details for such a research body. Because of the amazing diversity of this body of research, it is impossible to dive into the details of all 132 studies that we cover here. However, this article is targeted to general researchers in the fields such as information retrieval, natural language processing and ontologies. It is meant to give such researchers an introduction to the work in their related areas that have used Wikipedia with the goal of highlighting the potential of this rich corpus for their own work. Many studies have been conducted on various aspects of Wikipedia including its technical infrastructure, content, and contributors [93]. Researchers have examined Wikipedia's evolution over the years in terms of content and community. They investigated the coverage, quality, reliability, and readability of Wikipedia's content [80], and explored its readers and readership behaviors [94]. One of the most thorough literature reviews of Wikipedia is that of Medelyan et al. [78], that focuses specifically on the same subset of Wikipedia research that we address here: research that extracts and makes use of the concepts, relations, facts and descriptions found in Wikipedia, and organizes the work into four broad categories: applying Wikipedia to *natural language processing*; using it to facilitate *information retrieval* and *information extraction*; and as a resource for *ontology building*" [78, p.716]. Like this present review, they reviewed research that rather than studying Wikipedia itself as a phenomenon, uses the products of Wikipedia as a textual corpus for conducting text and multimedia oriented research. However, it is worthy to note that the size of Wikipedia has increased exponentially since their review and so did the scholarly articles that used Wikipedia as a a textual corpus.

Medelyan et al.'s review begins with a detailed description of the technical characteristics of Wikipedia's textual structure (such as articles, categories, and intra-wiki links) that facilitate corpus-oriented research. Their description is an invaluable introduction to Wikipedia for researchers, as they specifically focused on highlighting the features and characteristics that are most amenable to research analysis. The main sections of their review provide an in-depth examination of specific studies grouped by the topic of their research questions. In fact, we borrow our main categorization of the articles we present here from the structure of their review. However, we complement their main categories with much more detailed sub-categories providing a rich picture of what the main categories are about. An additional contribution of our review is a detailed trend analysis of the research details of the reviewed studies.

Because of their detailed coverage of past work, we generally do not duplicate the description of any article examined by Medelyan et al. Instead, we often identify these articles and refer readers to Medelyan et al. [78] for their descriptions. However, we analyzed all the corpus articles to extract the research details such as specific topics, research methodologies, and so on; thus, they are fully included in our analysis in the WikiLit website, which we explain later in this article. The articles we summarize and describe in this review are mostly those that were not included by Medelyan et al. (mainly because they were published after their review).

This review extends the existing literature by the following contributions.

- First, we summarize the principal means by which Wikipedia corpora have been used, extended, and enriched by other knowledge bases with the goal of excavating knowledge from large textual and multimedia data.
- Second, we trace numerous research details of the summarized studies and tabulate them to analyze the various approaches that have been adopted for researching corpus data.
- Third, we collect and categorize a number of tools that have been used to extract texts and images from Wikipedia in various formats.
- Finally, we further categorize a collection of disparately formatted datasets extracted from Wikipedia.

This review is not intended to identify Wikipedia's contributions to the reviewed topics. It is rather tailored to provide insights on various means for using Wikipedia content to enrich and/or test the methods and techniques developed in these topics. The summaries and research trends of the reviewed studies, the list of tools and datasets, provide significant resources for future research that intend to exploit the data available in Wikipedia. This article would be most helpful to researchers in the fields of information retrieval, natural language processing, and ontology building who are interested in Wikipedia's potential as an extensive natural-language corpus both in diversity of topics and in size for a wide range of research questions.

The rest of this paper is organized as follows. Section 2 summarizes the reviewed studies and analyzes the corpus research trends. These studies reside within three main areas closely tied with processing large data to present it semantically, uncover its hidden patterns and convert it into intelligent information; information retrieval (IR), natural language processing (NLP), and ontology building (OB). Next, a list of tools identified from the examined studies and elsewhere are described in Section 3. Another list of datasets is described in Section 4. Finally, the review concludes in Section 5.

2. Findings from Scholarly Research on Wikipedia

This review is part of a larger systematic review of scholarly research on Wikipedia. To ensure thoroughness and consistency in our review, we followed the guidelines presented by Okoli and Schabram [97]. These guidelines were designed to ensure the rigor of the methodology followed when conducting a systematic literature review. We also developed a plan in the form of a review protocol to assure consistency across the team members. This protocol included the selection process of studies to be included in the review [95]. It also detailed the data extraction process which answers a set of research questions extracted from each of the examined studies. Further details about the systematic review methodology are specified in a separate paper [93]. The selected studies are categorized according to their use of Wikipedia. We refer to these categories as the topics which encompass the following; general, infrastructure, content, participation, readership, and corpus.

The corpus category, which is the topic of this paper, includes 132 of the total number of identified studies. These studies do not examine Wikipedia directly, but rather use the content of Wikipedia as a data corpus for studying some other research questions.

Table 1: Corpus categories and number of studies in each subcategory

Corpus	132
Information Retrieval	62
Textual information retrieval	5
Multimedia information retrieval	4
Geographic information retrieval	3
Cross-language information retrieval	6
Data mining	5
Query processing	8
Ranking and clustering systems	15
Text classification	10
Other information retrieval topics	8
Natural language processing	46
Computational linguistics	6
Information extraction	17
Semantic relatedness	17
Other natural language processing topics	8
Ontology building	21
Other corpus topics	9

The corpus category comprises three main topics each of which is further divided into a number of sub-topics. Table I displays the categorization schema of the corpus topics and the number of studies in each category. The numbers of studies in the subcategories do not add up to 132 as some studies belong to more than one subcategory.

To complement the synthesis of the studies presented in this review, we built the WikiLit website (http://wikilit.referata.com). WikiLit offers a structured presentation of the categorization discussed above and the parsed details of each of the summarized articles. This website covers all the studies from all the categories including the corpus studies reviewed in this paper.

2.1. Use of Wikipedia as a Corpus

The corpus category, with 132 articles, discusses research using Wikipedia as a textual corpus for various text analysis studies. What is distinctive about this review is that the goals or outcomes of these studies are usually not focused on Wikipedia itself; they usually use Wikipedia content (both direct article text and metadata) as a textual data source for some other scientific analysis. We checked Wikimedia-pedia for a corresponding topic group but it was not available. Thus, we divided this topic category into four subcategories, three of which are obtained from [78] which we discussed earlier. In these three sub-categories, we identified research that used Wikipedia's content, including articles, hyperlinks, and statistical data for developing new methods, frameworks, techniques and systems, within three major areas: information retrieval (IR), natural language processing (NLP), and ontology building (OB). Our fourth subcategory comprises corpus topics that do not fit neatly into the other three.

2.1.1. Information Retrieval

Information retrieval (IR) is a broad area of study that aims to build systematic approaches to solve various challenges related to providing information search and access. The enormous collection of articles available in Wikipedia has encouraged IR researchers to use corpora (the plural of corpus) extracted from Wikipedia. Among the IR topics, we identified the textual or multimedia retrieval, information extraction, text classification, query processing, and data mining. The majority of the articles we found developed new methods or algorithms to enhance the performance of IR systems in terms of the relevance of the information retrieved and the query execution time.

Textual Information Retrieval. Studies in this section aim to improve the major IR tasks focused on text. These include query processing, computing relevance feedback and employing disambiguation techniques using Wikipedia.

In his dissertation, Evgeniy Gabrilovich [34] used linguistic information from Wikipedia and the Open Directory Project to improve text categorization performance. He used feature generation techniques to empower the training instances with more informative and discriminating features [34, p.7]. These features, which are concepts extracted from Wikipedia articles, are intended to enrich and complement the commonly used bag-of-words representation of documents with knowledge-rich features. For instance, a name of a public figure that appears in a text might be augmented by the figure's title, position, and other personal details. The infoboxes available in Wikipedia are very suitable resources for such information. In another doctoral thesis, Liu [75] modeled a new IR system by designing and incorporating a word sense disambiguation algorithm and expanding queries using Wikipedia and WordNet dictionaries. A relevance feedback assisted by Wikipedia was also employed to enhance the performance of the proposed IR system. The experimental results showed, in comparison to variations of Collins parser, an increase in the performance of the proposed system in terms of recall, precision, mean and geometric mean average precisions.

Bast et al. [8] presented ESTER, an efficient search engine that works based on a combination of full text and ontology search. The links between Wikipedia's articles were used to train a semi-supervised method that learns semantic information. Using the English version of Wikipedia and the YAGO ontology, ESTER was capable of efficiently processing various complex queries. In a similar line of research, Vechtomova [126] proposed a four-stage approach to retrieve blog posts that contain opinions about entities expressed in queries such as persons, products, or events [126, p.71]. After collecting and pre-processing the posts in the first stage, a number of faceted queries (disjunctions of a list of short queries) are built using Wikipedia in the second one. Afterwards, a topic-based document retrieval and opinion-based re-ranking methods are applied to maximize the relevancy of retrieved documents. This study demonstrated the importance of using Wikipedia for the identification of concepts that relate to opinion targets. Clark et al. [19] the development and evolution of genre in Wikipedia for the purpose of retrieving structured texts. By understanding how human categorize texts, a machine can be trained for automatic retrieval.

Wikipedia's textual content is continuously edited and added by a large number of contributors around the world. This rich textual repository provides an invaluable asset to research in the textual information retrieval field. Feature extraction, word disambiguation and query expansion are examples of applications that are enhanced

using Wikipedia.

Multimedia Information Retrieval. Multimedia databases, including images and videos available online, have exponentially increased, raising the need for new techniques to search these large collections. Wikipedia with both its growing multimedia contents was a target application for the studies included in this section.

Ah-Pine et al. [3] investigated multimedia information access. They proposed two novel approaches for hybrid text-image information processing that can be readily applied to the more general multimodal scenarios. They extended the principle of trans-media feedback into a metric view. The new similarity measures of cross-content provides the capability of finding expressive images for a text, to annotate an image, cluster or retrieve multi-modal objects. The authors used a dump of French Wikipedia images to evaluate their approaches. Rahurkar et al. [110] proposed a two-component application for image interpretation. The first component is responsible for keyword disambiguation using the titles of Wikipedia articles. The second one consists of an image to semantic concept mapping which is achieved by extracting semantic knowledge from Wikipedia. An image sorting system was developed based on the previous approach and an image sorting algorithm. Another popular application is image tagging that enables the filtering of large collections of images to retrieve specific ones. A useful tag for images is the location in which each of them was taken. Kalantidis et al. [59] proposed a new application, VIRal, for finding the location where a photo is taken using its visual content and Wikipedia geo-referenced articles. Using VIRal and an image as a query, the system returns visually similar images and estimates its location. To recognize the location, the authors used the points-of-interests and landmarks databases from Wikipedia. VIRal was challenged with a one million urban image dataset and proved efficient.

Beside texts and images, a large collection of videos is publicly available on the web. The classification of these videos can assist the video search and retrieval tasks. Perea-Ortega et al. [103] used Wikipedia's articles and Google searches to add more informational sources to assist the classification task of video data. VideoCLEF 2008 dataset and several supervised machine learning algorithms were used in various experiments to prove the enhancement of the video classification results using the web content. The machine learning algorithms include naive-Bayes, k-nearest-neighbors, support vector machines, and decision trees.

The increasing popularity of social networking web sites contributes immensely to the number of images published online. This raises the need for systems that are capable of efficiently organizing and filtering large collections of images. We foresee more research in this area of multimedia information retrieval in the years to come, in which Wikipedia could prove valuable.

Geographic Information Retrieval. Geographic Information Retrieval (GIR) extends the IR task by associating a geographic location feature to documents. Some Wikipedia articles have markup with geographical coordinates that can be extracted and used with rendered maps such as in Google Earth and the Danish Findvej.dk. The studies included in this section proposed different methods to solve the GIR task using the geospatial information available from Wikipedia.

Quack et al. [109] described an approach for mining images available on the web using unsupervised learning. The proposed system starts with a pool of geo-tagged

images from Flickr and a grid of geospatial tiles to build clusters of mined objects and events. These clusters are then assigned, via data mining techniques, text labels that are then employed to assign, in an unsupervised manner, each cluster to a Wikipedia article. These assignments are finally evaluated using the contents (texts and images) of the corresponding Wikipedia articles. Inaccurate assignments are filtered out and the final set of verified clusters-Wikipedia articles are then used to to geo-locate new images that have no geotags. Overell and Ruger [98] used Wikipedia corpus to generate co-occurrence models for place names disambiguation. Particularly, they used a rule-based method to annotate how places occur in Wikipedia and in what order. The choice of Wikipedia helped solving the synonyms issue by recording how multiple anchor texts are linked to the same Wikipedia article. The experimental results showed that the proposed approach enhanced the performance of GIR systems in terms of their mean average precision. Furthermore, using the Wikipedia corpus, Stokes et al. [117] investigated the success of natural language processing approaches to GIR tasks. They found that a careful choice of weighting schemes in the IR engine can minimize the negative impact of severe errors like toponym detection errors, toponym resolution errors, and query overloading. The authors employed the geospatial information available from Wikipedia to enhance the performance of their proposed system. Particularly noteworthy in Stokes et al. [117] study is their use of the geonames web service, which uses Wikipedia articles to describe geographic locations including countries, cities, colleges, universities, cultural icons, etc.

Wikipedia's presence in the GIR community through the context of GeoCLEF and its pilot track, GikiP is also noteworthy. Various GIR systems participated in this pilot track to answer questions with geographic reasoning by returning relevant articles from Wikipedia. Examples of such questions include which Dutch painters became famous by their portraits? Which European physicists immigrated to the US between the two world wars? These questions were posed in three different languages; English, German, and Portuguese. About $5,756,424$ of geo-referenced Wikipedia entries, written in 266 different languages, are accessible from Geonames and Wikipedia Webservice [1]. Entries that belong to the same category, such as "country" or "city" provide the same information through fixed-formatted tables called infoboxes. This structured information is pragmatic and advantageous in solving the geographic information retrieval problem.

Cross-language IR. Studies on cross-language IR (CLIR) used Wikipedia to improve the task of retrieving information in a language different from that of the user query. An example of such usage is WikiWord, a system that extracts lexical and semantic information from Wikipedia to build a multilingual thesaurus ([62], [63]). Wikipedia's inter-language links provide a rich tool to improve CLIR. Erdmann et al. [33] used these links to extract bilingual terminology. They implemented a Support Vector Machine (SVM) classifier, trained with a manually labeled data of term pairs and tested the performance of other extracted terms. Despite the promising results reported, the manual labeling of the data is quit expensive and time consuming. This raises the need for unsupervised methods to build a bilingual dictionary.

Lin et al. [74] and Lin et al. [73] described a Japanese-Chinese cross language IR system which is composed of four components; segmentation, translation, disambigua-

[1] http://www.geonames.org/wikipedia/

tion, and retrieval and re-ranking. The translation component consists of a Japanese-Chinese bilingual dictionary and Wikipedia inter-language links to translate query terms. To enhance Korean-Chinese IR (KCIR) tasks, Wang et al. [130] suggested a hybrid named entities translation from Korean to Chinese. The proposed system uses Wikipedia inter-language links as a translation tool to expand the bilingual dictionary and learns translation patterns directly from Google search results. The results of the experiments showed an improved KCIR performance in comparison with another method that only uses an offline dictionary. Potthast et al. [108] surveyed and compared the models for cross-language plagiarism detection dealing with analysis of similarities between texts from different languages. The evaluation of their experiments is performed on two corpora, Wikipedia corpus and RC-Acquis corpus. The authors reported the performance of three retrieval models; cross-language alignment-based similarity analysis (CL-ASA), cross-language explicit semantic analysis (CL-ESA), and cross-language character n-gram model (CL-CNG). It was found that "CL-CNG outperforms CL-ESA and CL-ASA" [108, p.15].

Another study that examined cross-language question answering using Wikipedia [36] is covered in [78]. Based on the studies above, Wikipedia has been employed in the four main categories of CLIR techniques; dictionary-based, parallel corpora based, comparable corpora based, and machine translator based CLIR techniques. The inter-language links is the main characteristic of Wikipedia exploited in such techniques.

Data Mining. Data mining, also referred to as data or knowledge discovery, is basically the process of extracting patterns from a large dataset. Extracted information aims to provide additional knowledge through these discovered patterns. This is also another main task in IR that motivated researchers to use Wikipedia as a data source to develop new mining systems. Different approaches were proposed in the studies covered in this section to mining information from large knowledge sources including Wikipedia.

In his dissertation, Zhang [142] proposed a new graph-based text mining system. A collection of texts was first represented using a graph and then enhanced using an ontology map or Wikipedia categories. Then, the structure of the graph with its nodes and edges was analyzed to uncover patterns to be used to enhance text clustering. Zhang also studied the effect of different types of linkage on text clustering. The graph-based methods presented herein were tested with two applications in the biomedical literature context: text clustering and summarization. The use of Wikipedia ontology was analyzed and compared to other methods when applied in text clustering systems. It was shown that the Wikipedia's category-based structure improve clustering results more than concept information [142, p.109].

Denoyer and Gallinari [29] described the XML Mining Track at INEX 2008. The aim of this track was to explore the classification and clustering of XML documents. A number of categorization and clustering models participated in this track. For testing purposes, the authors used a corpus of $100,000$ Wikipedia XML documents, their internal structure and the link information between documents. The recall of each of the categorization models as well as the macro-purity and micro-purity of each of the clustering models were reported.

To overcome the limitations of negative selection-based anomaly detection techniques in sparse data cases, Pöllä and Honkela [104] proposed a combination of symbol frequency analysis and negative selection. Wikipedia was employed as a real-world data to evaluate

the sensitivity of the proposed anomaly detection algorithm. In particular, the experiment analyzed the Wikipedia edit detection and showed promising results. Medelyan et al. [78] also described other data mining studies using Wikipedia ([9]; [85]).

The articles summarized here show that Wikipedia is a suitable corpus for solving a variety of tasks in the data mining field. These tasks include text classification and clustering, anomaly detection, and text summarization. Modeling the dependency between entities is another common data mining task that could be investigated using the internal and external links available from Wikipedia.

Query Processing. In this section, we summarize studies that aimed to expand queries dynamically by mining Wikipedia. The common objective of these studies is to enhance the relevancy of a query's results and its execution time.

Milne et al. [87] presented and discussed a new search interface called Koru. To understand the subject of both queries and documents, Koru derived a thesaurus for each document collection from Wikipedia. Wikipedia's articles were then used to model the building blocks of the thesaurus. The wiki and its hyperlinks were used to determine the connections between the thesaurus blocks. Elsas et al. [32] explored the blog feed retrieval task from two viewpoints; retrieval models and query expansion algorithms. The models developed in this study emphasized the importance of modeling the topical relationship between the feed and its entries. Moreover, a Wikipedia link-based query expansion method for feed retrieval proved to outperform other methods with no query expansion. Theobald et al. [123] described TopX, a system that intends to merge two points of view for processing top-k query for semi-structured data, database systems and IR. TopX's components are categorized as data-entry time or query-processing time. At the former, the documents' contents are indexed and the concepts and semantic relations are identified. At the latter, queries are decomposed and query keywords are mapped into available concepts. The Wikipedia corpus was used as a test bed and results showed that TopX was more effective and efficient than existing systems.

Machine learning techniques have also been employed in the query processing tasks of query classification and segmentation. Hu et al. [55] used Wikipedia articles and categories to solve the challenges of the query intent classification problem. Compared to other machine learning approaches, this method decreases the human effort to train a query intent classifier and improves the classification accuracy. Wikipedia knowledge base was also used by [122] to augment their proposed unsupervised learning approach to query segmentation. The use of Wikipedia as well as the Expectation-Maximization (EM) algorithm to optimize the proposed approach showed 46% improvement in comparison with other segmentation methods. Furthermore, Hwang [56] used the full English Wikipedia dataset exported in October 2007 to test and support the performance claims of a dynamic authority-based searching system. They proposed an approximation of the ObjectRank algorithm which materializes small subsets of the entire data graph. This helps reduce the query execution time as the algorithm needs to run only on one of the generated sub-graphs.

Chu [18] proposed new approaches for handling sparse relational datasets, specifically data extracted from unstructured documents. Chu addressed the RDBMS issues in handling sparse data, beginning by the construction of a workbench to extract and query structured from unstructured data. Various tools were then provided to query and process data. The new way of processing data stems from the "pay as you go" con-

cept which helps processing the data incrementally. Experiments to examine structured queries over Wikipedia were designed to test the performance of the workbench. Results showed that users were able to establish sophisticated queries. Chu also argued that his approach significantly eased the transition from extracting attributes from documents to querying them.

Kasneci et al. [60] proposed the NAGA query engine for the YAGO ontology described by Suchanek et al. [120]. YAGO facts are based on Wikipedia infoboxes and category names. YAGO-NAGA is proposed to extract information for building large scale knowledge bases. This is an ongoing work of maintaining and extending YAGO and providing a toolkit to extract information from it. The usefulness of YAGO has been demonstrated by its usage by various knowledge management projects such as DBpedia, SUMO, and UMBEL.

Relevancy and execution time are two main quality attributes of query retrieval systems. Researchers have long tried to optimize such systems to improve their quality, especially when dealing with large amount of data. Hence, the large number of Wikipedia articles presents a challenging dataset for such tasks. The Question Answering using Wikipedia pilot [2] is a project aimed at accessing Wikipedia content to answer specific queries. The choice of Wikipedia documents for this task was justified by Wikipedia being "one of the largest reference works ever, making it a natural target for question answering systems" [58]. Querying large databases continues to be a challenging task despite the availability of high performance servers and clusters. The extensive corpora available from Wikipedia, with its various forms of structured and semi-structured data, offer capable test beds for current and future query processing systems.

Ranking and Clustering Systems. Wikipedia has also been used to enrich texts from various sources to improve the performance and accuracy of ranking and clustering processes. In the context of ranking systems, the following studies presented various approaches to improve the performance of the ranking task.

To help narrow down the retrieved results of search engines, Gollapudi and Sharma [45] proposed a set of axioms for result diversification which can be viewed as a re-ranking process for the search results. The disambiguation pages in Wikipedia were used as an evaluation dataset to test the presented methods. The titles of these pages were used to draw ambiguous queries which are keyed in a search engine. The results of each search are then used to test the diversification algorithm. Hwang [56] proposed BinRank, an optimized dynamic link-based ranking method that approximates scores more efficiently than ObjectRank and PageRank. The core of this dissertation consists of providing efficient search functionality that enhances the usefulness of graph-structured data sources. This is achieved by taking advantage of the hidden information lying behind the semantic links available in such data sources. BinRank overcomes the limitation of traditional IR techniques in providing search results that are semantically meaningful. Wikipedia articles were used as a dataset to test the performance of BinRank since it contains semantic links such as 'definition' links, 'see also' links, and 'category' links. Using materialized views, BinRank approximates ObjectRank results achieving a faster and higher quality search results. Furthermore, Hwang et al. [57] used BinRank to present

[2] WiQA, http://ilps.science.uva.nl/WiQA/

a solution for a dynamic authority-based ranking problem by computing a number of materialized subgraphs. The full English Wikipedia data set exported in October 2007 was employed to test and support the performance claims of this solution.

Similarly, Lizorkin et al. [76] proposed an iterative method to compute the similarity between objects. This is based on SimRank, a graph-based measure that represents the relationships among objects by logical graphs. Intuitively, two objects that are connected to similar objects are considered to be similar. This approach was tested on a subset of the English Wikipedia including its articles and category links. An optimized version of the SimRank computation algorithm was also introduced and shown to converge 50 times faster the baseline model. Another ranking approach is built around the Google matrix which is formed by augmenting the normalized adjacency matrix with an extra term. The first eigenvector associated with the Google matrix determines the PageRank of an article. The adjacency matrix may be transposed, normalized and augmented. Its first eigenvector may be found to yield what Zhirov et al. [143] called the CheiRank. The authors used CheiRank and PageRank to "analyze the properties of two-dimensional ranking of all Wikipedia English articles and show that it gives their reliable classification with rich and nontrivial features" [143, p.523].

Wikipedia researchers have also examined the quantitative characteristics of the networks inherent in Wikipedia. Networks can be represented in matrices which, in context of Wikipedia, can be constructed from the content and metadata of Wikipedia articles. Mathematical operations can be performed on the constructed matrices to examine aspects of Wikipedia or to test computational algorithms on large-scale data. Buntine and Valtonen [12] built a matrix from the within-wiki links between $500,000$ pages of the English 2005 Wikipedia and used a discrete version of the hubs and authority algorithm to find topics in Wikipedia. For example, one topic would display the Wikipedia articles "Scientific classification" and "Animal" as the top authorities and "Arterial hypertension" and "List of biology topics" as the top hubs. They proposed the use of such link analysis models to specify an authority ranking for returning the results of search engines.

The two following studies focused on named entity ranking rather than ranking search results. Wikipedia provides a large assortment of defined entities such as "Location" and "Person". Zaragoza et al. [138] examined the problem of ranking entities of different heterogeneous sets of types by comparing two approaches, namely, the entity containment graphs and web search methods. They employed a statistical entity recognition algorithm to develop a snapshot of a semantically annotated English Wikipedia. Two observations from this study are noteworthy; first, the notion of inverted entity frequency is important to discount general types in entity containment graphs. Second, the rank of the documents in the computation of correlations enhanced the performance of the web search methods. Pehcevski et al. [102] implemented a new approach for entity ranking systems using the categories and link structure of Wikipedia. This approach also presented a topic classification based on extracted features from an INEX topic definition. The experiments conducted using the 2006 Wikipedia XML Corpus illustrated the advantages of using the categories and semi-structured data of Wikipedia to increase the effectiveness of entity ranking systems.

In an automated link discovery procedure, tools may suggest intrawiki links from a word in a Wikipedia article to an appropriate wiki page. For example, Adafre and Rijke [1] proposed a method to discover missing links in Wikipedia's articles by identifying a cluster of similar articles based on their titles and the co-citation information.

Candidate links that might be missing in the target article are then identified from some of the similar articles. Bai et al. [5] presented the Supervised Semantic Indexing (SSI) approach, that assigns a ranking score to documents based on their relevance to a query. The proposed methods were tested with Wikipedia documents taking advantage of Wikipedia's links structure. The experimental results proved that SSI performs better than vector-space models when applied on an Internet advertising data. In the clustering context, Banerjee et al. [7] improved the classification task of short texts such as blog feeds by extending each feed with additional features extracted from the titles of related Wikipedia articles. The results of the experiments proved that this approach improves the clustering accuracy [7, p.788]. The Wikipedia knowledge base was used by Carmel et al. [15] to enhance cluster labeling. Their approach begins by extracting a number of representative terms from the texts of a number of documents. Then, these terms are used to query Wikipedia and get the pages relevant to the corresponding cluster of documents. The meta-data of the returned pages such as the titles and the categories are used to label the cluster. For subjects that are well covered by Wikipedia, this method proved to assign very good clusters labels.

In addition, an important aspect of query retrieval systems is the organization of the search results into a hierarchy of labeled clusters. Carpineto et al. [16] proposed a new approach to solve the mobile search problem by grouping the search results into a set of labeled clusters. A set of queries known as ambiguous Wikipedia entries was used to analyze the subtopic relevance of search results. Nielsen [90] used non-negative matrix factorization in a hierarchical mode to cluster Wikipedia articles and scientific journals based on the scientific citations in Wikipedia. His algorithm identified scientific areas such as "cancer" and "immunology", each associated with a set of Wikipedia articles and a set of scientific journals.

Instead of working with Wikipedia links, the words within its articles may also be used as features in the construction of a matrix which can be viewed as a document-term matrix. A decomposition of such a matrix is often termed latent semantic analysis, particularly if a singular value decomposition method is used. For assessing the performance of newly developed algorithms, Řehůřek [127] constructed a document-term matrix from the entire English Wikipedia with the resulting size of 100,000 times 3,199,665, corresponding to a truncated vocabulary on 100,000 words and almost 3.2 million Wikipedia articles.

In addition to the above articles, another study that treated ranking and clustering systems is summarized in the Semantic Relatedness section of this review ([54]). With the exponential increase of information available on the Web, ranking the search engines' results is an arduous task. Moreover, many applications require the clustering of documents based on common themes or topics such as the grouping of search engine results into meaningful categories. The semantic knowledge embedded within the structural components of Wikipedia makes it exploitable for both tasks.

Text Classification. Text classification is a common problem in IR systems in which a classifier is trained to assign documents to appropriate classes. Studies in this section examined various methods to solve this problem benefiting from the large collection of documents available from Wikipedia. Several studies used Wikipedia knowledge base to enhance the text classification task.

Wang and Domeniconi [128] used Wikipedia to improve document classification by defining concept-based kernels. The representation of documents is augmented by extracted knowledge from Wikipedia in a semantic kernel form. The proposed approach works in both supervised and unsupervised learning settings. In other words, it works even if class labels of documents are not available. Testing this approach with four different datasets such as Reuters-21578 and OHSUMED showed better accuracy results than the bag of words (BOW) techniques. Similarly, Wang et al. [129] extended the BOW method with a thesaurus derived from Wikipedia to improve the text classification task. We summarize this study in the section on Semantic Relatedness.

Overell et al. [99] utilized Wikipedia's categories and templates as two structural patterns to extend the WordNet lexicon and develop a system, for classifying tags, ClassTag. The first component of the system classifies Wikipedia articles based on these structural patterns and lexicon. Tags are then mapped into the resultant categories in the second component. Two measures, recall and precision, were separately optimized to test the efficiency of ClassTag. Results showed improved performance when compared to standard WordNet. Considering the importance of metadata for the reusability of learning resources, Meyer et al. [81] compared the use of Wikipedia with that of traditional corpora for classification methods. They found that the use of Wikipedia successfully determined general topics, specific topics and subtopics of learning resources. This study also stressed the significance of using Wikipedia's categories and templates structures to improve the performance of text classification methods.

The remaining studies we discuss in this category used various methods in different classification applications. Farhoodi et al. [35] presented an automatic web page classification method, which they tested with the Persian Wikipedia. They demonstrated the pertinence of content-based and context-based features extracted from web pages in the proposed classification method. Murugeshan et al. [89] presented a profile based method for Wikipedia XML document classification, using negative category information. Ray et al. [111] discussed automatic question classification, a module of a question answering system. They proposed an answer validation solution using various resources including Wikipedia to validate answers offered by "open-domain Question Answering Systems" [111, p.1935]. Xiang et al. [136] proposed new approaches for text analysis and retrieval to address the gap between different knowledge areas and transfer the knowledge from one domain to another. The authors studied the knowledge gap between two domains and proposed a criteria to sample data from Wikipedia to fill this gap and then train a transductive SVM classifier on the augmented dataset.

In addition to the above articles, another study that dealt with text classification is summarized elsewhere in this review [2], and Medelyan et al. [78] described another [40]. The semantic information that can be extracted from Wikipedia enriches the traditional methods used for text classification, such as the Bag of Words. The semantic relationships between the terms in a text provide additional features that have proven to enhance the accuracy of text classification methods. Therefore, regardless of the features used to represent a text, Wikipedia corpora could be exploited to improve this representation.

Other Information Retrieval Topics. In addition to the IR topics described above, there are some articles concerning retrieving data or information from Wikipedia that do not fall under any of the labeled IR topics. Zhou et al. [144], for example, attempted to solve one of the main challenges in peer-to-peer file sharing systems, namely the supporting

content-based search. They proposed an adaptive indexing approach that consists of grouping and classifying terms based on their significance. This approach decreases the indexing cost and answers queries without the need for global knowledge. They validated their approach on various benchmark and Wikipedia datasets.

To accomplish a folksonomy visualization, Lee et al. [68] proposed a statistical model based on the frequency of each tag in Wikipedia articles to derive subsumption relations between tags. The neighboring tags were used to disambiguate the sense of a tag, since one word can be associated with multiple articles in Wikipedia. This method was tested with the del.icio.us tags and results proved its capability in visualizing the relationships between tags. Krizhanovsky and Smirnov [65] proposed a method for indexing wiki texts and tested it with Russian, English, and German Wikipedias. The limitations of the indexing method include the need for updating the index continuously. Also, the tf-idf weights do not consider the topics of the processed texts. Wikipedia's categories might be used to refine these weights. Pak and Chung [100] proposed a new method that improves the relevance of contextual ads using Wikipedia's articles as baselines to select and display ads on specific pages. The authors attributed their choice of Wikipedia to its wide coverage of a large variety of concepts as well as its updated content due to the frequent edits. A 50% improvement in the average matching precision was obtained using this method.

In his doctoral dissertation, Simma [116] developed different techniques to model the time distributions of event-driven data. These techniques are based on introducing for each event a Poisson process of its followers. These techniques were then tested with Tweets and the revision history of Wikipedia. The latter consisted of 414,540 pages with 71,073,739 revisions categorized as minor fixes, major insert, major delete, major change, revert and other edit. The experiments showed promising results in exploiting various delay, transition and fertility distributions. Friedlin and McDonald [38] investigated the medical knowledge of Wikipedia and used it to improve a laboratory and clinical observation database (LOINC). They found the medical knowledge of Wikipedia very extensive and useful as a scientific medical informatics resource, and the software they developed could satisfactorily add descriptions from Wikipedia articles to LOINC part names.

In addition to these studies listed here, Medelyan et al. [78] discussed other studies on information retrieval that we do not repeat here [86, 135, 134, 51, 4, 23].

2.1.2. Natural Language Processing

The ambiguous nature of natural language raises the need for computational linguistic analysis for the processing of languages in a range of applications. Natural language processing (NLP) can be applied to the translation of a text into another language, paraphrasing a text, and answering questions about the content of a text. Being a multilingual online encyclopedia, Wikipedia offers NLP researchers a semantically rich dataset. In this area, Wikipedia has been mainly employed for studies on computational linguistics and in semantic relatedness, as well as some other NLP topics.

Computational Linguistics. Computational Linguistics is a subfield of NLP that aims to derive functions to investigate and evaluate various facts about human language. In the introduction, we mentioned Gurevych and Wolf's presentation of lexical semantic

information from Wikipedia which is particularly valuable for computational linguistics studies. A wide variety of other studies have also investigated this topic.

One example of the ambiguous characteristic of human languages is polysemy; the ability of a word to have multiple meanings in different contexts. Word sense disambiguation is then the process of identifying, within a specific context, the most appropriate meaning of a polysemic word. In a highly cited study, [82] described a sense-tagging method that employs Wikipedia to identify the sense annotations. What makes Wikipedia a particularly suitable source of sense annotations is the disambiguation pages that contain links to the different meanings of an entry. Also, one of Wikipedia's editing norm is to link a word (or group of words) to the corresponding Wikipedia articles. Therefore, the proposed method consists of extracting all the paragraphs that contain an ambiguous word as part of a link or piped link which are used to extract the word's senses. The experiments demonstrated the reliability of the sense annotations extracted from Wikipedia and their suitability to construct accurate sense classifiers.

In another highly-cited study, [140] developed java-based APIs to extract information from Wikipedia and Wiktionary (JWPL and JWKTL, respectively). They also provided useful tools to support NLP studies for both academia and industry. Ganter and Strube [43] described a system for detecting linguistic hedges using Wikipedia weasel tags. Linguistic hedges are words or sounds that denote uncertainty in language. For example, the word "around" is a hedge in the following sentence: "I will get home at around 5 pm". Weasel tags are words or phrases that sound meaningful but in fact are vague; "People say that ..." The proposed system is based on words preceding weasels and added syntactic patterns. The experimental results demonstrate "that the syntactic patterns work better when using a broader notion of hedging tested on manual annotations" [43, p.176].

Turdakov and Kuznetsov [124] discussed several problems in word sense disambiguation and described available algorithms and methods used to solve them. They examined the method used in the Texterra system, which employs the Wikipedia corpus, and compared it to other methods in the literature. They argued for Wikipedia's suitability for such methods because of its structured document network and varied types of pages such as disambiguation and redirection pages. Furbach et al. [39] described LogAnswer, a German language system that answers questions using Wikipedia's knowledge base derived by computational linguistics and automated reasoning means. A semantic network representation of a snapshot of the German Wikipedia and $12,000$ logical rules were used as the knowledge base of the system. An additional article that examines other aspects of computational linguistics is discussed in the Introduction Section [48].

Leveraging large-scale knowledge base corpora has contributed to the progress in NLP solutions. An open problem in computational linguistics is word-sense disambiguation which Wikipedia attempts to solve using its "Disambiguation"' page. As such, a word with multiple meanings will have links to articles that describe each of its meanings. This is an additional advantage of using Wikipedia to improve the performance of computational linguistics algorithms.

Semantic Relatedness. Computing semantic relatedness among a set of documents or terms is a challenging task which assigns a similarity value based on the semantic content of these documents. The studies grouped in this section used the Wikipedia knowledge base to compute the semantic relatedness of words and documents.

Schenkel et al. [113] exploited Wikipedia's categories structure to add semantics to XML data. They generated an XML corpus (YAWN) from a 2006 Wikipedia dump using a Wiki2XML that converts Wiki markup to XML. Then, the constructed corpus was annotated with semantic tags that are extracted from Wikipedia's categories and lists, and WordNet. The authors suggest using this corpus in various applications including context-based information retrieval and Wikipedia pages clustering and classification. Turdakov and Velikhov [125] presented a similarity measure based on Dice's measure to compute the semantic relatedness between Wikipedia articles. Two articles are considered to be related if their Dice measure is high. This measure is computed as the ratio of the number of links the two articles have in common to the total number of links of both articles.

Semantic relations have been shown to enhance the performance of clustering algorithms. Hu et al. [54] built a concept thesaurus on the semantic relations extracted from Wikipedia to be used in a new text clustering method. Compared to traditional text clustering methods based on "bag of words", this method showed an enhanced clustering performance when applied with Reuters and OHSUMED datasets. In a similar study, Wang et al. [129] developed an automatic thesaurus of concepts from Wikipedia to enrich the"bag of words" representation of texts. This thesaurus aimed to capture the semantic relations between the words of a text to improve the text classification results. Several experiments were conducted using three different datasets; Reuters, 20NG, and OHSUMED. The classification performance of the proposed approach was measured using precision-recall metrics. The results showed the effectiveness of the added thesaurus.

Zesch et al. [141] and Zesch and Gurevych [139] investigated the literature to develop measures for computing semantic relatedness of word pairs. The identified measures were categorized into four types: path based, information content based, gloss based, and vector based measures. They compared these measures in relation with the datasets used (such as WorldNet or Wikipedia), the measure type, and the language (English or German). They concluded that crowds-based resources are not better than linguists-based resources. A higher precision but lower recall can be obtained by using the first paragraph of Wikipedia articles rather than the whole articles. In addition, this study presented two freely available systems, namely, DEXTRACT and JWPL. The former assists the construction of semantic relatedness datasets, and the latter provides a Java-based Wikipedia API for building NLP applications [139, p.25].

Pantel et al. [101] proposed using distributional and entity set expansion to improve the computation task of the semantic term similarities. The authors evaluated their approach using a collection of entity sets extracted from Wikipedia which they also made public to serve as a testbed for set expansion analysis. Experiments with 500 million terms underlined the effect of the corpus size and quality of the set expansion performance. They showed that the pairwise similarity of theses terms were computed in 50 hours using 200 quad-core nodes [101, p.946]. Holloway et al. [52] analyzed the semantic structure of Wikipedia and the coverage of its content. Basic statistics for the articles and categories were computed from a 2005 Wikipedia dump. Categories were particularly mapped using the cosine pairwise similarity measure. The category network was visualized and color coded by last edit time and top ten most active authors. The main finding of this study is that, although the category structure of Wikipedia is constructed by varied people and bots with different motives, it is actually well developed and maintained. Gabrilovich and Markovitch [42] introduced a new method for representing natural lan-

guage semantics, Explicit Semantic Analysis (ESA), that represents the meaning of any text in terms of concepts based on Wikipedia articles. They argued that the main advantage of their contribution is in handling synonymy and polysemy. ESA was tested in the text categorization context. When compared with previous methods, ESA enhanced the assessment of semantic relatedness of words and texts.

Li et al. [71] proposed a new approach for keyphrase extraction using topic relevance and term association. They represented a document as a weighted graph which vertices correspond to selected terms from the document and the weights denote the semantic relatedness among these terms. The use of Wikipedia in this method was in the selection of keyphrase candidates and the computation of their semantic relatedness. Different algorithms were employed then to relate documents by their topics and compute the term association. Experimental results showed that the keyphrase extraction approach proposed in this paper outperforms other approaches.

We discuss a couple of other semantic relatedness studies in more detail elsewhere in this study. In Ontology Building, we discuss the YAGO system, proposed by Suchanek et al. [120], which extracted data from Wikipedia and combined it with WordNet. In Information Extraction, we dicuss Grineva et al. [46] computation of semantic information for a new competitive key terms extraction method. In addition to these, other studies that examined various aspects of computing semantic relatedness using Wikipedia, ([41]; [106]; [107]; [72]) are discussed by Medelyan et al. [78].

Many applications involve the computation of semantic relatedness of texts, especially to solve NLP problems such as summarizing and clustering texts. Wikipedia's links, categories and lists of similar articles provide an exploitable resource that assists the identification of semantic relations between words and textual documents.

Information Extraction. In this sub-category of IR, studies used Wikipedia to extract structured information. Documents used for information extraction include text, HTML and XML pages.

Named Entity Recognition. One of the fundamental tasks of information extraction (IE), named entity recognition (NER), deals with identifying named entities such as proper names, names of organizations, locations, dates, or genes from freeform text. NER often relies on a machine learning algorithm and an annotated dictionary (gazetteer).

Bunescu [11] also aimed to derive new IE techniques with higher performance than existing ones using NER, named entity disambiguation and relation extraction. For NER, he considered the correlations between candidates named entities. These correlations were captured using Relational Markov Networks. Named entity disambiguation was achieved by detecting matches between proper names and named entities compiled from Wikipedia. A ranking function was used to compute the similarity value between proper names and named entities. Extracting relations between pairs of entities was solved using two types of supervised learning; single and multiple instance learning.

Mika et al. [84] used a NER tool to semantically annotate Wikipedia corpus linking articles texts to its infoboxes. The resulting annotations were then linked to DBpedia to enrich its metadata. This mapping between the semantic annotations and DBpedia was employed to generate additional sentences which are used to improve the initial annotation task. Demartini et al. [27] proposed a formal model for describing and ranking entities to solve the problem of entity retrieval (ER). Wikipedia page links and categories

were employed for query-category assignments. Combined with other natural language processing techniques, the performed tests showed an improvement of the ER task.

Named entity recognizers are essential processing resources in the majority of NLP solutions. Existing named entity recognizers, such as the one implemented by the Stanford NLP group, train a system with corpora such as REUTERS MUC-6, MUC-7, and ACE. Wikipedia, with its large collection of articles about various entities including persons, locations, and organizations, provides another valuable training set for these resources.

Keyword Extraction. Keyword extraction refers to the task of extracting keywords from a document or collection of documents. These keywords help users to filter documents based on their interests. They also assist IR tasks such text categorization and clustering.

Csomai [21] proposed a new approach for automated keyword extraction and its application to the back-of-the-book indexing. The goal of this study is to solve the keyword extraction problem with fewer resources and higher performance. After examining various supervised and unsupervised keyphrase extraction techniques, Csomai found that keyphrase extraction can definitely be used to automate the back-of-the-book indexing task. The indexing process should be modularized where each module handles different stages of the process. Such modules include candidate extraction, phrase ranking and phrase filtering. In addition, a combination of different candidate extraction methods leads to better results than the state of the art tf-idf method. This dissertation also considered new features based on statistical measures and linguistic features based on semantic analysis. Mihalcea and Csomai [83] and Csomai and Mihalcea [22] proposed Wikify, a system for keyword extraction and word-sense disambiguation using Wikipedia. Wikify identifies prominent concepts in a document and links them to Wikipedia pages. The tests employed demonstrated Wikify's improvement in the time taken to answer questions.

Wikipedia was used by Grineva et al. [46] as a knowledge base to derive semantic information for a new competitive key terms extraction method. A document is first represented by a graph of semantic relationships among its terms. The dense part of the graph depicts the document's main topics while the sparse part represents the less important terms. Afterwards, the graph is partitioned using graph community detection techniques. A criterion function is then used to select groups with important terms. Wikipedia is utilized to extract information necessary to compute the terms weights and their semantic relatedness. The main advantages of this approach include the elimination of a training phase and the effectiveness with noisy and multi-theme documents.

Devereux et al. [30] investigated the challenges of the acquisition of feature-based conceptual modeling. They argued for the significance of three types of knowledge, namely, encyclopedic, syntactic and semantic in guiding the feature extraction process. They also proposed a new feature extraction method using class-based information. Two automatically parsed Wikipedia corpora were used to evaluate the proposed feature extraction method. Results were promising given the relatively small size of the employed corpora. The authors concluded that using the entire Wikipedia might improve the feature extraction performance.

Similar to named entity recognition, keyword extraction is an additional processing phase that serves in the majority of NLP tasks. The extracted keywords provide a labeling or classifying scheme for the corresponding texts. The general topic of a text or document is usually identified by its assigned keywords. The text corpora, links, and

infoboxes available from Wikipedia could be further exploited to enhance the performance of existing keyword extraction methods.

Open Information Extraction. Open IE differs from traditional IE in the scalability of their corresponding systems. For instance, the former requires homogeneous corpora and a fixed number of relations. However, the open IE scales to heterogeneous large corpora and a flexible number of relations. Weld et al. [132] explored the challenges and benefits of open IE in the context of Kylin. Kylin is an IE system that uses Wikipedia infoboxes to train relationally-targeted extractors using self-supervised learning. Kylin's goal is to help scaling to the Web the conversion of unstructured text to relational form. This study highlighted the importance of combining the relational approach used in Kylin with the different structural forms available in Wikipedia including its infoboxes, edit histories, and multi-lingual links, to potentially improve the precision and recall of open IE systems. Even though only one study is summarized in this section, the open IE category is significant enough to be a category of its own. Its flexibility in types of corpora admitted make open IE more favorable when the corpora is as large as the web.

Other NLP Topics. Many other studies explored different NLP applications using Wikipedia. Coursey [20] described a new machine learning algorithm, WikiRank, that is developed to assign values to each entry of an encyclopedic knowledge source. Based on links that associate the entries of an encyclopedia (in this case, Wikipedia), these assigned values can be used in various NLP applications: automatic topic identification, text-based paraphrases recognition, and ontology terms recognition.

Dorji et al. [31] presented a new method for Field Association (FA) terms that are words or phrases used to identify document fields in document classification. This method extracts FA terms using part-of-speech (POS) pattern rules and rank them using a modified version of the tf-idf weights. The authors evaluated their method using a 306 MB Wikipedia dump. Mehler et al. [79] utilized the complex network theory to develop an automatic language classification model. They tested their variant of the Sapir-Whorf Hypothesis using a corpus of 160 Wikipedia-based social ontologies. Stone et al. [118] used Wikipedia to compare different models for paragraph similarity analysis, and to automatically generate similar smaller corpora. When comparing single paragraphs, the results favored the use of simple models such as word overlap over more complex ones such as Topic Model and LSA. Ferschke et al. [37] presented the Wikipedia Revision Toolkit, an open source toolkit which is usually used with the Java Wikipedia Library (JWPL) we described earlier. The main features of this toolkit include the reconstruction of past states of Wikipedia and the access to all article revisions. Moreover, this toolkit provides a vital knowledge source based on Wikipedia's edit history to enhance the NLP processing algorithms. In addition to those discussed here, Medelyan et al. [78] described other NLP studies whose details we extracted on the WikiLit website, but that we do not summarize here [119, 105, 112, 23].

2.1.3. Ontology Building

Ontology, in the information science context, can be simply defined as the description of a set of concepts within a domain and of the relationships between those concepts. Ontology building (OB) has attracted the interest of a large number of researchers in the last decade especially with the exponential increase of available data online. Researchers

have recognized Wikipedia as a major data source to support their work in developing new web ontologies. The foremost reason for building ontologies is analyzing and enabling the reuse of domain knowledge. The articles grouped in this subcategory presented different approaches to OB using Wikipedia.

Hepp et al. [49] presented a framework that exploits wiki-based technology to build ontologies. As an example, they showed an application where Wikipedia entries were considered as ontology elements. Muchnik et al. [88] argued that content is not the only factor that dictates the hierarchy of concepts; context is also important. Directed networks of terms were employed to handle context. They proposed five different statistical methods designed to construct a hierarchy in networks of related terms. They also implemented a complex network analysis package that has been employed in various network research studies.

Given a number of ontologies, Kim et al. [61] proposed an approach to merging and matching ontologies based on two modules; a linguistic module to compute similarities between concepts and a topic map constraints-based module. In their experiments, the authors used the Wikipedia philosophy ontology, oriental philosophy ontologies, western philosophy ontologies, and the Yahoo western philosophy dictionary. The ontologies matching results agreed to a great extent with the manual matching performed by domain experts. To overcome the non-scalability limitation of existing ontology learning methods, Wong et al. [133] proposed a new clustering algorithm, called Tree-Traversing-Ant (TTA). They used the TTA algorithm along with two measures for term similarity and dissimilarity: normalized Google distance and number of Wikipedia which is based on the cross-linking of Wikipedia articles. Their empirical tests showed 48% ontological improvement. After building, merging or matching ontologies, it is crucial to have means for evaluating these ontologies. Therefore, Yu et al. [137] presented ROMEO, a requirement-oriented methodology for evaluating ontologies. ROMEO imposed five ontology requirements on Wikipedia. Since there is no strict hierarchy imposed on the Wikipedia category structure, the first requirement was to ensure a sufficient intersection level between categories. The second requirement provides a guideline on how to group categories adequately to guarantee that the category structure is useful and efficient in browsing articles. The third requirement stresses the fact that cycles should be avoided in the category structure as they can lead to users being lost in a cycle of navigation. The fourth ontology requirement is for ensuring the completeness of the available set of categories. The last requirement calls for the correctness of the set of associated categories associated, that is, no articles are incorrectly placed in multiple categories.

McCrae and Collier [77] presented a method that automatically generates regular expression patterns and develops a thesaurus. A classifier was trained using the BioCaster ontology from the biomedical domain to classify terms as synonymous or non-synonymous. The proposed method was compared with Wikipedia and WordNet and experiments showed promising performance. In the field of nucleic acid research, Gardner et al. [44] presented the Rfam database that archives non-coding ribonucleic acids (RNAs) using sequence alignments and statistical profile models. They discussed the pros and cons of using Wikipedia for community-driven annotation. They recommended Wikipedia for other curation methods while stressing that, because Wikipedia is built by consensus, a loss of tight control of the data allowed by in-house curation is to be expected.

Capocci et al. [14] compared imposed classifications to real clustering in a particular

case of a scale-free network [14, p.1], Wikipedia. The performed analysis showed that links in Wikipedia do not necessarily denote similarity, which encourages testing the clustering algorithm's performance with manual indenxig before employing it for automatic categorization. Guo et al. [47] proposed an ontology learning technique which relies on socially emergent bodies of knowledge like Wikipedia to build ontologies rather than the traditional expert knowledge. The resulting ontologies were comparable to traditional ones.

Lehmann et al. [69] developed an interactive visualization tool that exhibits the structure of visited articles in Wikipedia and highlights other relevant topics. The objective of their approach was to assist users in following possible paths in their searches while increasing the relevance of the obtained information. Tests of this approach demonstrated the viability of the proposed interface in browsing and searching Wikipedia. Hu [53] used Wikipedia to enrich ontologies with Wikimantics, vectors extracted from Wikipedia articles. Hu referred to these vectors as "Wikipedia-enhanced concept descriptors" [53, p.470]. Wikimantics were shown to be useful to several applications including ontology matching, with the limitation of being strongly tied to one repository, Wikipedia.

Cantador et al. [13] used Wikipedia categories as a semantic knowledge base for the purpose of transforming social tags into ontology concepts in the task of automatic tag categorization. Furthermore, Wikipedia entries and URIs were used to annotate web content. Each entry in the English version of Wikipedia is considered a unique identifier for the concept described in the corresponding entry, and so can be exploited as an ontology component. The approach they proposed was evaluated on a dataset collected from Flickr. The results showed the improvement achieved using content and context based tags instead of subjective and organizational ones.

Additionally, other studies considered Wikipedia to solve various real world problems. Banchuen [6] developed a geographical analogue engine that computes the similarity within textual information and combine the results with those of numeric methods used for characterizing places. Wikipedia articles were used to create an ontology using the Web Ontology Language (OWL) that computer algorithms can manipulate. Banchuen explored techniques from various fields including artificial intelligence, linguistics, cognitive science, and knowledge engineering. The experimental results highlighted several observations related to different semantic measures, such as the statistical description, template description, complete stop-words list, and complete vocabulary. Sigurdsson and Halling [114] used Wikipedia topics related to music for the MuZeeker search engine, grouping search results according to Wikipedia categories. Syed [121] proposed a knowledge base derived automatically from Wikipedia and other similar information sources that organize world knowledge in a standard machine readable format. This would allow computer applications to better access and exploit knowledge in different forms.

Finally, two major projects, YAGO and DBpedia, are very significant and noteworthy that we conclude this section with their summaries. YAGO is described by Suchanek et al. [120] as an ontology that is based on the concepts derived from Wikipedia infoboxes and the taxonomies available from WordNet. The evaluation of YAGO showed a 95% precision according to the type checking techniques they employed. YAGO has been exploited in various applications: semantic search, entity organization, information extraction and ontology construction. DBpedia is thoroughly described by Bizer et al. [10] as the project that produced one of the largest knowledge bases extracted from

Wikipedia. It contains the descriptions of "more than 3.64 million things out of which 1.83 million are classified in a consistent Ontology, including 764,000 persons, 573,000 places, 112,000 music albums, 72,000 films, 18,000 video games, 192,000 organisations, 202,000 species and 5,500 diseases" [17]. DBpedia can handle complex queries against Wikipedia via SPARQL query builders and interfaces. It also links other available online datasets to Wikipedia information. Among others, the British Broadcasting Corporation uses DBpedia for linking documents across their web site [64].

In addition to those discussed here, two other articles discuss other aspects of ontology building in Wikipedia. The first can be found elsewhere in this review [66] and the second, [4], is reviewed by [78].

Wikipedia, as a collaborative work of an enormous number of volunteers, has helped the move away from traditional approaches to ontology construction in which the source of knowledge emerged only from experts. In the Web 2.0 and Wikipedia age, researchers shifted towards collaborative bodies of knowledge as a source for building ontologies. Moreover, Wikipedia was also used to evaluate the ontology available in its category structure to support browsing in Wikipedia. Consequently, Wikipedia can benefit from its own structure to improve its content.

2.1.4. Other Corpus Topics

Letia et al. [70] addressed the design of a new commutative replicated data type (CRDT) algorithm, treedoc, to solve the consistency problem in large-scale systems. The CRDT aims to make concurrent updates commute. Wikipedia revision pages were stored as treedocs with each revision being the result of one of two operations; insert or delete. CRDT showed better performance compared to traditional approaches. Another CRDT algorithm, Logoot-Undo, was presented in Weiss et al. [131] with the undo anywhere, anytime feature which operates on highly dynamic content in P2P network. This algorithm was validated using the revisions of some Wikipedia pages. As such, the MediaWiki API was employed to dump these revisions to XML files. Then, using a diff algorithm the differences between two revisions were identified and only reverts modifications were finally considered. The experiments showed that Logoot-Undo "ensures the CCI (Causality, Convergence, and Intention) consistency model" and maintains a low overhead when applied with Wikipedia's corpus [131, p.1172]. Curino et al. [25] suggested new methods as part of a new system, PRISM, to solve the time-consuming and error-prone problems of the schema evolution task. "Continuous validation against challenging real-life evolution histories, such as the one of Wikipedia, proved invaluable in molding PRISM into a system that builds on the theoretical foundations laid by recent research and provides a practical solution to the difficult problems of schema evolution" [25, p.772]. Curino et al. [24] aimed to provide a deep analysis of the evolution of databases in Web Information Systems (WIS). For instance, the authors studied the evolution of Wikipedia database and schema. This study concluded by highlighting the need of tools of automation of documenting database and schema evolution especially in the case of WIS which are open and more dynamic.

Silva et al. [115] used Wikipedia content, specifically its articles to construct a network of mathematical theorems. The authors employed the diversity entropy method to identify the centrality of each theorem. According to their results, the oldest theorems tend to be the most important ones, in the sense that they have higher values of diversity

entropy, in the average. On the other hand, the frontiers theorems are those recently added to the network [115, p.6].

Denoyer and Gallinari [28] described a corpus compiled of articles from eight language Wikipedias converted to XML. The corpus consists of article pages and categorizations of these articles arranged in various useful configurations, and has proven extremely popular for information retrieval and ontological research. An additional article, David Ahn et al. [26], discusses the use of Wikipedia at the TREC QA track and is summarized by Medelyan et al. [78].

2.2. Corpus Research Trends

To further analyze the use of Wikipedia in scholarly articles, we categorize the reviewed studies by various research details. These details were parsed from each article and stored in our WikiLit website which provides interactive tools to analyze these details. The analysis we present in this section allows researchers to learn about various trends of corpus research.

2.2.1. Year

Even though Wikipedia was launched in January 2001, the first corpus related studies were identified in 2005. Two years later, the number of studies encountered its major lift reaching 24 articles in 2007. The drop of number of studies in 2011 is due to the fact that our search extended to only part of that year, as explained in the introduction. Table 2 shows that the majority of the studies are related to the IR topics followed by NLP and OB.

	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Information Retrieval							
Cross-language IR					4	2	
Data mining		1	1		3	1	
Geographic IR				3			
Multimedia IR					1	3	
Other IR topics				3	1	4	
Query processing			1	3	1	1	
Ranking and clustering systems	1		2	2	4	6	
Text classification		1		2	4	3	
Textual IR		2	1		1	1	
Natural Language Processing							
Computational linguistics			1	1	3		
Information extraction			6	7	4		
Other natural language processing topics		1	2		1	1	2
Semantic relatedness		1	6	4	5	1	
Ontology Building			7	5	4	4	1
Other corpus topics	1	2		2	1	3	
Total number of distinct studies	**2**	**8**	**24**	**31**	**31**	**33**	**3**

Table 2: Wikipedia Corpus studies by year

2.2.2. Wikipedia Coverage

Table 3 classifies the corpus studies according to the degree of their treatment of Wikipedia. As expected, given the focus of this literature review, 73 of the studies

treated Wikipedia as sample data in the course of studying some other primary focus, benefiting from the variously formatted corpora extracted from Wikipedia. 43 studies regarded Wikipedia as the main topic of investigation, whereas 8 studies only considered Wikipedia as a case among others.

	Case	**Main topic**	**Other**	**Sample data**
Information Retrieval				
Cross-language IR		1	5	
Data mining		3		3
Geographic IR		1	1	1
Multimedia IR		2	2	
Other IR topics			1	7
Query processing		2	1	3
Ranking and clustering systems		5		10
Text classification		3	2	5
Textual IR	1			4
Natural Language Processing				
Computational linguistics	2	3		1
Information extraction	1	7	3	6
Other natural language processing topics		1	1	6
Semantic relatedness	1	7	1	8
Ontology Building	3	7	4	7
Other corpus topics		4		5
Total number of distinct studies	**8**	**43**	**17**	**73**

Table 3: Wikipedia Corpus studies by their degree of focus on Wikipedia

2.2.3. Wikipedia Language

The corpus studies utilized different language versions of Wikipedia as displayed in Table 4. To save some space in the table, we used the first two letters of the languages' names. "All" refers to all languages, "MU" to multiple languages, and "NS" to no language specified. The majority of the studies (76 of 132) explicitly mentioned that they used the English version of Wikipedia. However, we can surmise that the 48 studies that did not explicitly specify the language also used the English version, as we only included studies published in English; this would give a total of 124 studies using the English (En) Wikipedia. 8 studies handled multiple language versions, 6 German (Ge), 4 Chinese (Ch), 3 French (Fr), and 2 Spanish (Sp). Additionally, one study was found to treat each of the following language versions: Dutch (Du), Korean (Ko), Persian (Pe), and Russian (Ru).

2.2.4. Unit of Analysis

As illustrated in Table 6, the great majority of the corpus studies exploited Wikipedia articles (Art) as their unit of analysis. This finding is not surprising given the nature of the IR, NLP and OB studies. The remaining studies, in decreasing order, examined the individual edit, website (Web), category (Cat), user, scholarly article (Sc.A) and subject (Sub) as their units of analysis.

	All	Ch	Du	En	Fr	Ge	Ja	ko	NS	MU	Pe	Ru	Sp
Information Retrieval													
Cross-language IR		3		3		1	2	1	1				1
Data mining				3					1	2			
Geographic IR	1			1						1			
Multimedia IR				1	1					2			
Other IR topics				4		1				4		1	
Query processing				4						2			
Ranking and clustering systems				11			1			4			
Text classification				4	1	1				4	1		
Textual IR				2						3			
Natural Language Processing													
Computational linguistics				2		2				3			
Information extraction			1	9					1	7			1
Other natural language processing topics				6					1	1			
Semantic relatedness		1		11		1				5			
Ontology Building	1			12	1				2	6			
Other corpus topics				3					2	4			
Total number of distinct studies	**2**	**4**	**1**	**76**	**3**	**6**	**3**	**1**	**8**	**48**	**1**	**1**	**2**

Table 4: Wikipedia Corpus studies by Wikipedia Language version

	Art	Cat	Edit	Sc.A	Sub	User	Web	N/A
Information Retrieval								
Cross-language IR	6							
Data mining	3		1				1	1
Geographic IR	3			3				
Multimedia IR	4							
Other IR topics	4		1		1			2
Query processing	2					1		3
Ranking and clustering systems	14							1
Text classification	10							
Textual IR	4	1						
Natural Language Processing								
Computational linguistics	3						1	2
Information extraction	15							2
Other natural language processing topics	6		1					1
Semantic relatedness	17		1			1		
Ontology Building	14	3				1	2	2
Other corpus topics	5		2	1			1	1
Total number of distinct studies	**110**	**4**	**6**	**1**	**1**	**3**	**5**	**15**

Table 5: Wikipedia Corpus studies by unit of analysis of Wikipedia data

2.2.5. Wikipedia Data Extraction

Another important trend of scholarly research on Wikipedia is the means by which data was extracted from Wikipedia. The findings related to this trend are displayed in Table 7. Whereas 78 studies extracted data from a cloned Wikipedia database, 51 studies used the live version of Wikipedia.

2.2.6. Wikipedia Page Type

Various Wikipedia page types were used in the corpus studies as shown in Table 8. The article page type was examined in 125 studies, 59 of which are IR related. The NLP based studies dealt mostly with the information categorization and navigation (ICN) page type. This is due to the inner links between the different language versions of Wikipedia articles.

	Clone	Live	Secondary	N/A
Information Retrieval				
Cross-language IR	2	4		
Data mining	2	3	1	
Geographic IR	1	2		
Multimedia IR	2	1	1	
Other IR topics	2	5		1
Query processing	6			
Ranking and clustering systems	10	4	1	
Text classification	8	2		
Textual IR	1	4		
Natural Language Processing				
Computational linguistics	4	1		1
Information extraction	9	6	1	1
Other natural language processing topics	4	3		1
Semantic relatedness	14	3		
Ontology Building	10	9	1	1
Other corpus topics	3	4	1	1
Total number of distinct studies	**78**	**51**	**6**	**6**

Table 6: Wikipedia Corpus studies by their data extraction from Wikipedia

	Article	History	ICN	Log	N/A	Other	Policy
Information Retrieval							
Cross-language IR	6						
Data mining	5				1		
Geographic IR	3						
Multimedia IR	4						
Other IR topics	6		1		2		
Query processing	5			2	1		
Ranking and clustering systems	15						
Text classification	10						
Textual IR	5						
Natural Language Processing							
Computational linguistics	5				1		
Information extraction	15				1	1	
Other natural language processing topics	7	1	1				
Semantic relatedness	17		3				
Ontology Building	16		2		3	1	1
Other corpus topics	6				3		
Total number of distinct studies	**125**	**1**	**7**	**2**	**12**	**2**	**1**

Table 7: Wikipedia Corpus studies by the type of Wikipedia pages treated

2.2.7. Research Design

For each of the studies examined in this review, we identified the research design utilized to analyze Wikipedia's corpora. The findings depicted in Table 9 reveal that the most commonly used approach is the experiment (EX) with 95 studies. This is conforming to the nature of IR, NLP and OB studies, which usually require an experiment to test the performance of their proposed approaches. Subsequently, the case study

(CS) and statistical analysis (SA) research designs were employed in 11 and 19 studies, respectively. The rest of the studies are distributed as follows; 13 applied mathematical modeling (MM), 3 conceptual (CP), 2 design science (DS), 2 content analysis (CA), and 1 in each of ethnography (ET), literature review (LR), action research (AR), and grounded theory (GT).

	AR	**CS**	**CP**	**CA**	**DS**	**ET**	**EX**	**GT**	**LR**	**MM**	**SA**	**Other**
Information Retrieval												
Cross-language IR							4			2		
Data mining		1					3				1	1
Geographic IR							2			1		
Multimedia IR							3			1	1	
Other IR topics							8			1		
Query processing		1				1	5				1	
Ranking and clustering systems							11			5	1	
Text classification							10					
Textual IR				1			4			1		
Natural Language Processing												
Computational linguistics		2					3					1
Information extraction	1	2	1				10	1			3	
Other natural language processing topics							5			1	2	
Semantic relatedness				1	1		14				2	
Ontology Building		4	2		1		10			1	6	1
Other corpus topics		1					4		1		2	1
Total number of distinct studies	**1**	**11**	**3**	**2**	**2**	**1**	**96**	**1**	**1**	**13**	**19**	**4**

Table 8: Wikipedia Corpus studies by research design

2.2.8. Collected Datatype

Table 9 displays the type of data collected in the reviewed studies. Only 5 studies were not empirical and are labeled as "N/A". Among the remaining studies, the great majority (118 studies) extracted data from various Wikipedia pages (WP). The main topics that used Wikipedia pages to collect data were the following; OB (15 studies), information extraction (14 studies), ranking and clustering systems (14 studies), and semantic relatedness (13 studies). The next common data types collected were archival records (AR) with 19 studies and websites (Web) other than Wikipedia with 16 studies. Few studies collected data from experiment responses (ER, 5 studies), documents (DC, 3 studies), direct observation (DO, 5 studies), literature review (LR, 2 studies), survey responses (SR, 1 study), Computer usage logs (CUL, 1 study), and Interviews (IN, 1 study).

2.2.9. Corpus Topics by Domain of Knowledge

Table 10 presents the reviewed studies by their domain of knowledge. As could be expected, computer science (CS) is the dominating field with 122 studies. The other interdisciplinary (ID) domains identified include, in decreasing number of studies, information science (IS, 11 studies), geography (GE, 4 studies), and health (HE, 3 studies). Fewer studies were identified in social sciences (SS, 7 studies), humanities (HU, 3 studies), logic and mathematics (MA, 1 study), and natural sciences (NS, 1 study). This shows that there is still room for more contributions and studies in these domains in which Wikipedia serves as a large collection of corpora.

	AR	CS	DO	DC	ER	IN	LR	SR	Web	WP	N/A
Information Retrieval											
Cross-language IR	2			2						5	1
Data mining	2						1		1	3	
Geographic IR									2	3	
Multimedia IR				1						4	
Other IR topics	2									8	
Query processing		1	2		1			1	1	4	
Ranking and clustering systems	2								2	14	
Text classification	2								1	10	
Textual IR									1	4	1
Natural Language Processing											
Computational linguistics					2					5	1
Information extraction	4					1				14	
Other natural language processing topics									2	8	
Semantic relatedness	4		1						1	13	1
Ontology Building	1				2				5	15	1
Other corpus topics							1			8	
Total number of distinct studies	**19**	**1**	**3**	**3**	**5**	**1**	**2**	**1**	**16**	**118**	**5**

Table 9: Wikipedia Corpus studies by collected datatype

	HU	IN	MA	NS	SS	CS	GE	HE	IS
Information Retrieval									
Cross-language IR		6				6			
Data mining		5			1	5			
Geographic IR		3				2	2		1
Multimedia IR		4				3	1		1
Other IR topics		8				8		1	
Query processing		6				6			
Ranking and clustering systems	1	14				13			1
Text classification		10				9			1
Textual IR		5				4			1
Natural Language Processing									
Computational linguistics		6				6			
Information extraction		17				17			
Other natural language processing topics		7			1	16			
Semantic relatedness		16			1	16			
Ontology Building	2	19			4	13	1	2	5
Other corpus topics		8	1	1		8			
Total number of distinct studies	**3**	**134**	**1**	**1**	**7**	**122**	**4**	**3**	**11**

Table 10: Wikipedia Corpus studies by their domain of knowledge

3. Tools to exploit Wikipedia corpora

The large amount of data available from Wikipedia has spurred researchers as well as practitioners to create tools to extract various types of data from Wikipedia. While examining the studies included in this review, we compiled a list of tools that were employed to extract data from Wikipedia. We also added other tools and datasets that we collected from various sources. We further classify these tools by the functions they provide. In another paper, we provide an extensive list of tools and datasets for Wikipedia research [Okoli et al. 2012]. However, here we only list the tools that are particularly valuable for corpus-focused research. Such tools are used to extract text and images from Wikipedia or to handle various semantic and linguistic functions. The list of identified tools is displayed in Table 11.

Tool Function	Tool Name	URL
Information Extraction	wikipedia2text	http://www.evanjones.ca/software/wikipedia2text.html
	WikipediaFS	http://en.wikipedia.org/wiki/WikipediaFS
	SONIVIS	http://sonivis.org/wiki/index.php/Prepared Database# Wikipedia based data sets
	infobox2rdf	http://code.google.com/p/infobox2rdf/
	JWPL	http://code.google.com/p/jwpl/
	WikiXRay	http://meta.wikimedia.org/wiki/WikiXRay
	WikiExtractor	http://medialab.di.unipi.it/wiki/index.php/Wikipedia Extractor
Image Extraction	Catdown	http://toolserver.org/platonides/catdown/catdown.php
	Image for Biographies	http://wikipapers.referata.com/wiki/ Images_for_biographies
Semantic and Linguistic	Wikipedia-Similarity	http://www.hits.org/english/research/nlp/download/ wikipediasimilarity.php
	Manypedia	http://www.manypedia.com/
	WikipediaMiner	http://wikipedia-miner.cms.waikato.ac.nz/

Table 11: Tools to extract data from Wikipedia

3.1. Tools for text extraction

- As part of a NLP course, Evan Jones developed the *wikipedia2text* [3] tool to extract text from Wikipedia. This command-line program downloads and formats a specified Wikipedia article then displays it on the command-line. A modification of this tool called "AtD" generates plain text from a complete Wikipedia dump [4].
- *WikipediaFS* makes raw text Wikipedia articles available under the Linux file system so that a Wikipedia article can be viewed and edited as real files that exist on the local hard drive.
- *SONIVIS* is a piece of software that extracts information from various wikis including Wikipedia based on social networks analysis.
- *infobox2rdf* is a tool that generates RDF datasets from the infobox data available in Wikipedia dump files.
- Java Wikipedia Library, *JWPL*, is a Java-based application programming interface (API) that provides access to all information in Wikipedia.
- *WikiXRay* is a tool developed by Jos Felipe Ortega. It is written in both Python and R and used to download and process data from the Wikimedia sites for generating graphics and data files with quantitative results.
- *WikiExtractor* is a tool developed in the Medialab at the University of Pisa, Italy. It is implemented in Python and used to extract cleaned text from Wikipedia dumps.

[3] http://www.evanjones.ca/software/wikipedia2text.htm
[4] http://blog.afterthedeadline.com/2009/12/04/generating-a-plain-text-corpus-from-wikipedia/

3.2. Tools for image extraction

- *Catdown* is a tool to download images from Wikipedia by their corresponding categories. Direct access and use of Catdown is available at http://toolserver.org/ platonides/catdown/catdown.php.
- Emilio J. Rodríguez-Posada developed a tool that suggests images for biographies in several Wikipedias.

3.3. Semantic and Linguistic tools

- *Wikipedia-Similarity* is a tool used to compute semantic similarity using Wikipedia [119, 105].
- *Manypedia* provides a comparison of Linguistic Points Of View (LPOV) of different language versions of Wikipedia.
- *WikipediaMiner* is a toolkit for tapping into the rich semantics encoded within Wikipedia.

4. Wikipedia-based datasets

We also identified all published datasets generated from Wikipedia and used in some of the reviewed studies. In addition, we identified other valuable datasets from other sources. Due to the large number of datasets, we only list the most common ones in Table 13.

Dataset usage	**Dataset Name**	**Dataset URL**
IR	Wikimedia Downloads	http://download.wikimedia.org/
	Wikimedia Foundation Image Dump	http://archive.org/details/wikimedia-image-dump-2005-11
	Koblenz Network Collection	http://konect.uni-koblenz.de/
	page-to-page link	http://haselgrove.id.au/wikipedia.htm
	Wikipedia3	http://labs.systemone.at/wikipedia3
	Wikipedia edit history	http://snap.stanford.edu/data/wiki-meta.html
	Tamil	https://github.com/tshrinivasan/tamil-wikipedia-word-list
	Wikipediadoc	http://www.searchdaimon.com/community/dataset
NLP	Wikicorpus	http://www.lsi.upc.edu/ nlp/wikicorpus/
	WikiTaxonomy	http://www.h-its.org/english/research/nlp/download/
	WikiNet	http://www.h-its.org/english/research/nlp/download/
	WikiRelations	http://www.h-its.org/english/research/nlp/download/
OB	DBpedia	http://wiki.dbpedia.org/Datasets
	WEX	http://wiki.freebase.com/wiki/WEX

Table 12: Wikipedia Datasets per category

4.1. Datasets for Information Retrieval

- *Wikimedia Downloads:* a corpus of compressed XML files of Wikipedia from its official database dumps.
- *Wikimedia Foundation Image Dump:* 296,000 archived images from Wikipedia and its related projects.
- *Koblenz Network Collection:* a large network of datasets of all types.
- *page-to-page link:* downloadable files that contain all links between 5,716,808 Wikipedia pages.
- *Wikipedia3:* a monthly updated conversion of the English Wikipedia into RDF.
- *Wikipedia edit history:* complete Wikipedia edit history until January 2008.
- *Tamil:* a word list extracted from the Tamil Wikipedia dump.
- *Wikipediadoc:* a collection of 67,537 Wikipedia articles converted to Microsoft Word 2002 .doc format.

4.2. Datasets for Natural Language Processing

- *Wikicorpus:* a corpus that contains large portions of Catalan, Spanish and English Wikipedia (based on a 2006 dump) enriched with linguistic information.
- *WikiTaxonomy:* a taxonomy extracted from Wikipedia categories.
- *WikiNet:* a multi-language ontology developed by exploiting various aspects of Wikipedia.
- *WikiRelations:* a dataset that contains binary relations obtained from processing Wikipedia category names and the category and page network.

4.3. Datasets for Ontology Building

- *DBpedia:* a large domain ontology derived from Wikipedia.
- *WEX:* or Freebase Wikipedia Extraction is a processed dump of the English Wikipedia in XML and tabular formats.

5. Discussion

The research trends showed an overall increase in the number of studies that used Wikipedia since its birth. Although it is very difficult to draw broad conclusions from such a vast body of research, in this section we highlight some of the outstanding strengths of Wikipedia that is evident from studies that utilized its corpus. Table 14 summarizes the Wikipedia pages used in each of the identified tasks and applications in the examined studies.

Textual Information Retrieval: The natural language expressed in texts is often vague, which challenges the effectiveness of textual IR systems. Wikipedia provides a large number of articles that can be exploited to generate new features to enrich the representation

Task/Application	Wikipedia Resources
Feature generation	Wikipedia articles and link structures
Query Expansion	Wikipedia articles
Relevance feedback	Wikipedia articles
Image retrieval	Wikipedia images
keyword disambiguation for image sorting	Titles of Wikipedia articles
Image geo-tagging	Wikipedia geo-referenced articles
Video classification	Wikipedia articles
Image mining Wikipedia	Geo-referenced articles
Place names disambiguation	Wikipedia articles
Bilingual terminology extraction	Wikipedia inter-language links
Translation	Wikipedia inter-language links
Text clustering	Wikipedia categories
Categorization of XML documents	Wikipedia XML documents
Anomaly detection	Wikipedia articles
Named entity disambiguation	Named entities compiled from Wikipedia
Semantic annotation	Wikipedia texts and infoboxes
Query-category assignments/ Entity retrieval	Wikipedia page links and categories
Semantic relatedness/Key terms extraction	Wikipedia articles
Keyword extraction	Wikipedia articles (linked to concepts in a text)
Open Information Extraction	Wikipedia infoboxes, edit histories, and multi-lingual links
Question Answering	Wikipedia articles
Thesaurus construction/Question processing	Wikipedia articles and links
Query segmentation	Wikipedia articles
Dynamic ranking	Wikipedia articles and links (definition links, 'see also' links, and category links)
Similarity ranking	Wikipedia articles and category links
Query segmentation	Wikipedia articles
Search results ranking	Wikipedia articles and links
Document clustering	Wikipedia categories and titles of Wikipedia articles (for labeling clusters)
Document classification	Wikipedia articles (to augment the representation of documents)
Tag classification	Wikipedia categories and templates (as ontology extensions)
Sense annotations	Wikipedia articles
Linguistic hedges detection	Wikipedia weasel tags
Semantic relatedness of word pairs	Wikipedia articles
Semantic relations extraction	Wikipedia articles
Explicit Semantic Analysis/Text categorization	Wikipedia articles
Keyphrase extraction	Wikipedia articles

Table 13: Wikipedia and Textual IR

of texts and help improving these systems. [40] proposed mapping fragments of texts to corresponding concepts (Wikipedia articles) based on similarity algorithms. The links between articles were also considered to augment the concepts related to a given text.

Information extraction: Wikipedia knowledge-base can be used to derive semantic information to enhance the performance of keyword extraction [46]. Specifically, Infoboxes, pages URIs, lists and categories can be used to improve the recall of information extraction methods [134]. Moreover, Wikipedia articles can be used to enrich terms and concepts from unstructured texts [86].

Ranking and clustering systems: Features extracted from texts and documents are vital in improving the accuracy of the clustering task. Therefore, Wikipedia serves as an additional database of extra features to be considered in the texts and documents clustering. For instance, Banerjee et al. [7] linked each of a blog's feeds to the title of

a corresponding Wikipedia article. This additional feature proved to improve the feeds clustering accuracy.

Text classification: Similar to the clustering task, the accuracy of the classification of texts and documents depends on the features extracted from these documents. Augmenting the modeling of documents by using Wikipedia related articles enriches the features used in the classification process. This approach was tested with popular datasets such as Reuters-21578 and OHSUMED and lead to higher accuracy than the traditional bag of words technique. Wikipedia categories provide a supplementary resource that assists the text classification task. These categories present a dynamic classification scheme that consists of a priori classified articles.

Semantic relatedness: 17 of the total 31 NLP studies focused on using Wikipedia to elevate the accuracy of the computation of the semantic relatedness between different terms, documents, and various concepts. The links between terms and articles in Wikipedia provide a rich data to compute the semantic similarity between different concepts. This network of interconnected semantics has proven to boost the performance of the semantic relatedness task[106].

Ontology building: Wikipedia categories are found to play multiple roles in the reviewed studies. Besides assisting the text classification task, they also supply a semantically structured knowledge base that can be leveraged to build, evaluate and match ontologies. DBpedia is the largest and most popular ontology built on concepts derived from Wikipedia infoboxes [4]. It incorporates about 2350000 instances distributed among 359 classes the most common of which are place, person, work, species, and organization. YAGO is another extensive ontology built from Wikipedia, WordNet, and GeoNames. These knowledge bases encourage further exploitation of Wikipedia structured data to solve other ontology related research questions such as matching, and merging ontologies.

Limitations. One of the most common criticisms of Wikipedia is its susceptibility to vandalism, since the vast majority of articles can be edited by literally any Internet user, with no need even for registration. In theory, vandalism could indeed pose a problem for the purity of corpora, since at the moment any article is captured for research analysis, there is no way to guarantee that no vandalism exists in that article. Fortunately, in practice, the risk of any particular article containing vandalism is extremely low because of the effectiveness of numerous software agents specifically created to detect vandalism, as well as the vigilance of several human editors who spend their time patrolling for vandalism. Thus, most incidences of vandalism last only for a few minutes [50]. For any research analysis that uses a large number of articles, the possible noise created by the rare incidences of vandalized text is probably so negligible as to have no statistically significant effect on results.

Another controversy is Wikipedia's credibility as a citable resource for academic research [80]. A quick search on this topic leads to a number of university and library pages that warn their students not to cite Wikipedia. Nevertheless, these same pages agree that Wikipedia could be used to collect background information about any topic, although careful scrutiny is required. The main argument against citing Wikipedia is the absence of a mandatory review process for any published material. Wikipedias contributors are not required to be experts in the fields they are writing about. Therefore, there is no guarantee that Wikipedia provides accurate information at all times especially when a page is newly created or edited. In fact, Wikipedia itself highlights that an "encyclopedia

is a starting point for research, not an ending point"[5]. It also emphasizes the importance of not treating information in Wikipedia as definite truth.

This review has shown that Wikipedias corpora were proven to be useful in an assortment of applications. These are mostly text-based except for a few applications such as image interpretation and geo-tagging. Given its strict guidelines when it comes to non-free and copyrighted content, Wikipedia is still poor in audio and video content.

Another limitation of Wikipedia is a result of its "neutral point of view" policy which is also the reason for its success. This policy opposes the inclusion of opinionated article content that could be used in a wide spectrum of NLP applications, including opinion mining and sentiment analysis. These applications aim at computing the polarity of textual records, but Wikipedia articles are supposed to be always neutral. For example, the majority of known brands have dedicated Wikipedia pages that state their history, corporate structure, and financial figures. Whereas companies might want to examine these pages for actionable insights that could enhance theirs products, services, and marketing strategies, the required objectivity of the Wikipedia content means that these pages cannot be used to gain subjective insights. This being said, although this might be a limitation from the perspective of market research, it is consistent with Wikipedia's role as an encyclopedia, which is supposed to maintain objectivity in the topics it covers.

6. Conclusion

Wikipedia, with its rich and large content, has motivated many studies from different areas. In this review, we examined 132 studies from three research areas; IR, NLP, and OB. 91 of these studies were published as peer-reviewed journals and doctoral theses, and the remaining 41 are highly cited conference papers. These studies were summarized and analyzed to extract insightful trends of research related to the use of Wikipedia as a corpus. Moreover, this review identified and described a variety of tools used to extract data from Wikipedia as well as datasets of Wikipedia content. Regardless of the debatable aspects of Wikipedia content, this review highlights the value of Wikipedia in solving challenging applications that require large datasets. Its large corpora have been practically employed to ameliorate the performance of popular and impactful solutions for challenging problems such as information extraction, textual classification, and semantic relatedness. Nevertheless, we believe that Wikipedia's geospatial information and interlanguage links could be further exploited to solve more geographic and cross-language IR problems. The continuous growth of Wikipedia will continue to be attractive for researchers to mine and exploit its enormous corpora with its unstructured as well as semantically-structured components.

7. Acknowledgments

We want to note that all five co-authors were intensely involved in this project, and each one of us spent hundreds of hours on its execution. We thank Weiwei Zhang for her assistance in verifying the accuracy of research details of the WikiLit studies. We

[5] https://en.wikipedia.org/wiki/Wikipedia:Academic_use

thank Kira Schabram for her invaluable assistance in developing the systematic literature review methodology used [97], and in conducting the pilot study [96]. We thank Bilal Abdul Kader for his assistance in the pilot study [92]. We thank Richard Wong for his assistance in collecting author data. We thank Emilio J. Rodrguez-Posada (emijrp) for his model WikiPapers site, upon which much of WikiLit was based. We thank the innumerable researchers on the wiki-research-l and authors of included studies for their many comments and revisions. Earlier versions of this study have been previously published in a conference [92] and as a working paper [91]. The protocol for this study was presented in a conference [96], published as a working paper [95], and discussed in a workshop [67]. This study was funded by the Social Sciences and Humanities Research Council of Canada; the Lundbeck Foundation Center for Integrated Molecular Brain Imaging (CIMBI); the Concordia University Aid to Scholarly Activity fund; and the Danish Council for Strategic Research through the Responsible Business in the Blogosphere project.

8. References

[1] Adafre, S. F., Rijke, M. d., 2005. Discovering missing links in wikipedia. In: Proceedings of the 3rd international workshop on Link discovery. New York, NY, USA, pp. 90 – 97.

[2] Adar, E., Skinner, M., Weld, D. S., 2009. Information arbitrage across multi-lingual wikipedia. In: 2nd ACM International Conference on Web Search and Data Mining, WSDM'09, February 9, 2009 - February 12, 2009. Barcelona, Spain, pp. 94–103.

[3] Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., Renders, J.-M., 2009. Crossing textual and visual content in different application scenarios. Multimedia Tools and Applications 42 (1), 31–56.

[4] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. DBpedia: a nucleus for a web of open data. In: 6th International Semantic Web Conference, ISWC 2007 and 2nd Asian Semantic Web Conference, ASWC 2007, November 11, 2007 - November 15, 2007. Vol. 4825 LNCS. Busan, Korea, Republic of, pp. 722–735.

[5] Bai, B., Weston, J., Grangier, D., Collobert, R., Sadamasa, K., Qi, Y., Chapelle, O., Weinberger, K., 2010. Learning to rank with (a lot of) word features. Information Retrieval 13 (3), 291–314.

[6] Banchuen, T., 2008. The geographical analog engine: Hybrid numeric and semantic similarity measures for U.S. cities. Ph.D. thesis, The Pennsylvania State University, United States – Pennsylvania.

[7] Banerjee, S., Ramanathan, K., Gupta, A., 2007. Clustering short texts using wikipedia. In: 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07, July 23, 2007 - July 27, 2007. Amsterdam, Netherlands, pp. 787–788.

[8] Bast, H., Chitea, A., Suchanek, F. M., Weber, I., 2007. ESTER: efficient search on text, entities, and relations. In: 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07, July 23, 2007 - July 27, 2007. Amsterdam, Netherlands, pp. 671–678.

[9] Bhole, A., Fortuna, B., Grobelnik, M., Mladenic, D., 2007. Extracting named entities and relating them over time based on wikipedia. Informatica (Ljubljana) 31 (4), 463–468.

[10] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S., 2009. DBpedia - a crystallization point for the web of data. Journal of Web Semantics 7 (3), 154–165.

[11] Bunescu, R., 2007. Learning for information extraction: From named entity recognition and disambiguation to relation extraction. Ph.D. thesis, The University of Texas at Austin, United States – Texas.

[12] Buntine, W., Valtonen, K., 2005. Topic-specific scoring of documents with discrete PCA. In: ICML 2005 Workshop 4: Learning in Web Search. pp. 34–41.

[13] Cantador, I., Konstas, I., Jose, J. M., Mar. 2011. Categorising social tags to improve folksonomy-based recommendations. Web Semantics: Science, Services and Agents on the World Wide Web 9 (1), 1–15.

[14] Capocci, A., Rao, F., Caldarelli, G., Jan. 2008. Taxonomy and clustering in collaborative systems: the case of the on-line encyclopedia wikipedia. Europhysics Letters 81 (2), 28006–1.

[15] Carmel, D., Roitman, H., Zwerdling, N., 2009. Enhancing cluster labeling using wikipedia. In: 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, July 19, 2009 - July 23, 2009. Boston, MA, United states, pp. 139–146.
[16] Carpineto, C., Mizzaro, S., Romano, G., Snidero, M., 2009. Mobile information retrieval with search results clustering: Prototypes and evaluations. Journal of the American Society for Information Science and Technology 60 (5), 877–895.
[17] ChrisBizer, August 2012. Dbpedia 3.8 released, including enlarged ontology and additional localized versions.
URL http://blog.dbpedia.org/
[18] Chu, E., 2008. Sparse relational data sets: Issues and an application. Ph.D. thesis, The University of Wisconsin - Madison, United States – Wisconsin.
[19] Clark, M., Ruthven, I., Holt, P. O., 2009. The evolution of genre in wikipedia. Journal for Language Technology and Computational Linguistics 24 (1), 1"22.
[20] Coursey, K., 2009. The value of everything: Ranking and association with encyclopedic knowledge. Ph.D. thesis, University of North Texas, United States – Texas.
[21] Csomai, A., 2008. Keywords in the mist: Automated keyword extraction for very large documents and back of the book indexing. Ph.D. thesis, University of North Texas, United States – Texas.
[22] Csomai, A., Mihalcea, R., Oct. 2008. Linking documents to encyclopedic knowledge. IEEE Intelligent Systems 23 (5), 34–41.
[23] Cucerzan, S., 2007. Large-scale named entity disambiguation based on wikipedia data. In: Proc. 2007 Joint Conference on EMNLP and CNLL. Vol. 6. Prague, Czech Republic, p. 708"716.
[24] Curino, C. A., Moon, H. J., Tanca, L., Zaniolo, C., Jun. 2008. Schema evolution in wikipedia - toward a web information system benchmark. In: ICEIS 2008 - 10th International Conference on Enterprise Information Systems, June 12, 2008 - June 16, 2008. Vol. DISI. Barcelona, Spain, pp. 323–332.
[25] Curino, C. A., Moon, H. J., Zaniolo, C., 2008. Graceful database schema evolution: the PRISM workbench. In: Proceedings of the VLDB Endowment VLDB Endowment Hompage. Vol. 1. pp. Volume 1 Issue 1, August 2008.
[26] David Ahn, Valentin Jijkoun, Gilad Mishne, Karin Muller, de Rijke, M., Stefan Schlobach, 2005. Using wikipedia at the TREC QA track. In: The Thirteenth Text Retrieval Conference (TREC 2004).
[27] Demartini, G., Firan, C., Iofciu, T., Krestel, R., Nejdl, W., Oct. 2010. Why finding entities in wikipedia is difficult, sometimes. Information Retrieval 13 (5), 534.
[28] Denoyer, L., Gallinari, P., 2006. The wikipedia XML corpus. SIGIR Forum 40 (1), 64 – 9.
[29] Denoyer, L., Gallinari, P., 2009. Overview of the INEX 2008 XML mining track. Advances in Focused Retrieval, Jaap Kamps Archives and Information Studies/Humanities, University of Amsterdam, Amsterdam, The Netherlands 1012 XT.
[30] Devereux, B., Pilkington, N., Poibeau, T., Korhonen, A., 2009. Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. Research on Language and Computation 7 (2-4), 137 – 170.
[31] Dorji, T., Atlam, E.-s., Yata, S., Fuketa, M., Morita, K., Aoe, J.-i., Apr. 2010. Extraction, selection and ranking of field association (FA) terms from domain-specific corpora for building a comprehensive FA terms dictionary. Knowledge and Information Systems, 1–21.
[32] Elsas, J. L., Arguello, J., Callan, J., Carbonell, J. G., 2008. Retrieval and feedback models for blog feed search. In: 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM SIGIR 2008, July 20, 2008 - July 24, 2008. Singapore, Singapore, pp. 347–354.
[33] Erdmann, M., Nakayama, K., Hara, T., Nishio, S., 2009. Improving the extraction of bilingual terminology from wikipedia. ACM Transactions on Multimedia Computing, Communications and Applications 5 (4).
[34] Evgeniy Gabrilovich, Dec. 2006. Feature generation for textual information retrieval using world knowledge. Doctoral dissertation, Technion " Israel Institute of Technology, Haifa, Israel.
[35] Farhoodi, M., Yari, A., Mahmoudi, M., Oct. 2009. A persian web page classifier applying a combination of content-based and context-based features. International Journal of Information Studies 1 (4), 263–71.
[36] Ferrandez, S., Toral, A., Ferrandez, O., Ferrandez, A., Munoz, R., 2009. Exploiting wikipedia and EuroWordNet to solve cross-lingual question answering. Information Sciences 179 (20), 3473–3488.
[37] Ferschke, O., Zesch, T., Gurevych, I., 2011. Wikipedia revision toolkit: efficiently accessing wikipedia's edit history. In: Proceedings of the 49th Annual Meeting of the Association for Com-

putational Linguistics: Human Language Technologies: Systems Demonstrations. Stroudsburg, PA, USA, p. 97"102.
[38] Friedlin, J., McDonald, C. J., May 2010. An evaluation of medical knowledge contained in wikipedia and its use in the LOINC database. Journal of the American Medical Informatics Association: JAMIA 17 (3), 283–287.
[39] Furbach, U., Glckner, I., Helbig, H., Pelzer, B., Apr. 2010. Logic-based question answering. KI - Knstliche Intelligenz 24 (1), 51–55.
[40] Gabrilovich, E., Markovitch, S., 2006. Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence. Vol. vol.2. Menlo Park, California, p. 1301"1306.
[41] Gabrilovich, E., Markovitch, S., 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on Artifical intelligence. pp. 1606–1611.
[42] Gabrilovich, E., Markovitch, S., 2009. Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research 34, 443–498.
[43] Ganter, V., Strube, M., 2009. Finding hedges by chasing weasels: hedge detection using wikipedia tags and shallow linguistic features. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. pp. 173–176.
[44] Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn, R. D., Nawrocki, E. P., Kolbe, D. L., Eddy, S. R., Bateman, A., Nov. 2010. Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res., gkq1129.
[45] Gollapudi, S., Sharma, A., 2009. An axiomatic approach for result diversification. In: Proceedings of the 18th international conference on World wide web.
[46] Grineva, M., Grinev, M., Lizorkin, D., 2009. Extracting key terms from noisy and multitheme documents. In: Proceedings of the 18th international conference on World wide web. p. Wolfgang Nejdl L3S and Hannover University.
[47] Guo, T., Schwartz, D., Burstein, F., Linger, H., Sep. 2009. Codifying collaborative knowledge: using wikipedia as a basis for automated ontology learning. Knowledge Management Research & Practice 7 (3), 206–17.
[48] Gurevych, I., Wolf, E., Nov. 2010. Expert-built and collaboratively constructed lexical semantic resources. Language and Linguistics Compass 4 (11), 1074–1090.
[49] Hepp, M., Siorpaes, K., Bachlechner, D., Oct. 2007. Harvesting wiki consensus: using wikipedia entries as vocabulary for knowledge management. IEEE Internet Computing 11 (5), 54–65.
[50] Hicks, J., February 2014. machine kills trolls: How wikipedia's robots and cyborgs snuff out vandalism. http://www.theverge.com/2014/2/18/5412636/this-machine-kills-trolls-how-wikipedia-robots-snuff-out-vandalism, accessed: 2016-06-28.
[51] Hoffman, R., 2008. A wiki for the life sciences where authorship matters. Nature Genetics 40, 1047"1051.
[52] Holloway, T., Bozicevic, M., Borner, K., Jan. 2007. Analyzing and visualizing the semantic coverage of wikipedia and its authors. Complexity 12 (3), 30–40.
[53] Hu, B., Dec. 2010. WiKi'mantics: interpreting ontologies with WikipediA. Knowledge and Information Systems 25 (3), 445.
[54] Hu, J., Fang, L., Cao, Y., Zeng, H.-J., Li, H., Yang, Q., Chen, Z., 2008. Enhancing text clustering by leveraging wikipedia semantics. In: 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM SIGIR 2008, July 20, 2008 - July 24, 2008. Singapore, Singapore, pp. 179–186.
[55] Hu, J., Wang, G., Lochovsky, F., Sun, J.-t., Chen, Z., 2009. Understanding user's query intent with wikipedia. In: Proceedings of the 18th international conference on World wide web. p. Wolfgang Nejdl L3S and Hannover University.
[56] Hwang, H., 2010. Dynamic link-based ranking over large-scale graph-structured data. Ph.D. thesis, University of California, San Diego, United States – California.
[57] Hwang, H., Balmin, A., Reinwald, B., Nijkamp, E., 2010. BinRank: scaling dynamic authority-based search using materialized subgraphs. IEEE Transactions on Knowledge and Data Engineering 22 (8), 1176–1190.
[58] Jijkoun, V., Rijke, M., 2007. Overview of the WiQA task at CLEF 2006. Evaluation of Multilingual and Multi-modal Information Retrieval, Springer–Verlag Berlin, Heidelberg 2007.
[59] Kalantidis, Y., Tolias, G., Avrithis, Y., Phinikettos, M., Spyrou, E., Mylonas, P., Kollias, S., Nov. 2010. VIRaL: visual image retrieval and localization. Multimedia Tools and Applications, 1–38.
[60] Kasneci, G., Ramanath, M., Suchanek, F. M., Weikum, G., 2008. The YAGO-NAGA approach to

knowledge discovery. SIGMOD Record 37 (4), 41–47.
[61] Kim, J.-M., Shin, H., Kim, H.-J., 2007. Schema and constraints-based matching and merging of topic maps. Information processing & management 43 (4), 930–945.
[62] Kinzler, D., 2008. Automatischer aufbau eines multilingualen thesaurus durch extraktion semantischer und lexikalischer relationen aus der wikipedia. Diplomarbeit an der abteilung fr automatische sprachverarbeitung, Institut fr Informatik, Universitt Leipzig.
[63] Kinzler, D., 2009. WikiWord: multilingual image search and more. In: Wikimania.
[64] 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 connections. In: The Semantic Web: Research and Applications. Vol. 5554. Berlin/Heidelberg, p. 723"737.
[65] Krizhanovsky, A., Smirnov, A., 2009. On the problem of wiki texts indexing. Journal of Computer and Systems Sciences International 48 (4), 616–624.
[66] Krtzsch, M., Vrandecic, D., Volkel, M., Haller, H., Studer, R., 2007. Semantic wikipedia. Web Semantics 5 (4), 251–261.
[67] Lanamäki, A., Okoli, C., Mehdi, M., Mesgari, M., Aug. 2011. Protocol for systematic mapping of wikipedia studies. In: Proceedings of IRIS 2011 " The 34th Information Systems Research Seminar in Scandinavia. Turku, Finland, pp. 458–459.
[68] Lee, K., Kim, H., Jang, C., Kim, H.-J., Jun. 2008. FolksoViz: a subsumption-based folksonomy visualization using the wikipedia. Journal of KISS: Computing Practices 14 (4), 401–11.
[69] Lehmann, S., Schwanecke, U., Dorner, R., 2010. Interactive visualization for opportunistic exploration of large document collections. Information Systems 35 (2), 260–269.
[70] Letia, M., Preguica, N., Shapiro, M., 2010. Consistency without concurrency control in large, dynamic systems 44, 29–34.
[71] Li, D., Li, S., Li, W., Gu, C., Li, Y., 2010. Keyphrase extraction based on topic relevance and term association. Journal of Information and Computational Science 7 (1), 293–299.
[72] Li, Y., Huang, K., Ren, F., Zhong, Y., 2008. Searching and computing for vocabularies with semantic correlations from chinese wikipedia.
[73] Lin, C.-C., Wang, Y.-C., Tsai, R. T.-H., 2010. Japanese-chinese information retrieval with an iterative weighting scheme. Journal of information science and engineering 26 (2), 685–697.
[74] Lin, C.-C., Wang, Y.-C., Yeh, C.-H., Tsai, W.-C., Tsai, R. T.-H., 2009. Learning weights for translation candidates in japanese-chinese information retrieval. Expert Systems with Applications 36 (4), 7695–7699.
[75] Liu, S., 2006. Improve text retrieval effectiveness and robustness. Ph.D. thesis, University of Illinois at Chicago, United States – Illinois.
[76] Lizorkin, D., Velikhov, P., Grinev, M., Turdakov, D., 2010. Accuracy estimate and optimization techniques for SimRank computation. VLDB Journal 19 (1), 45–66.
[77] McCrae, J., Collier, N., 2008. Synonym set extraction from the biomedical literature by lexical pattern discovery. BMC Bioinformatics 9 (1), 159.
[78] Medelyan, O., Milne, D., Legg, C., Witten, I. H., 2009. Mining meaning from wikipedia. International Journal of Human Computer Studies 67 (9), 716–754.
[79] Mehler, A., Pustylnikov, O., Diewald, N., 2010. Geography of social ontologies: Testing a variant of the sapir-whorf hypothesis in the context of wikipedia.
[80] Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., Lanamäki, A., 2015. The sum of all human knowledge: A systematic review of scholarly research on the content of wikipedia. Journal of the Association for Information Science and Technology 66 (2).
[81] Meyer, M., Rensing, C., Steinmetz, R., 2008. Using community-generated contents as a substitute corpus for metadata generation. International Journal of Advanced Media and Communication 2 (1), 59–72.
[82] Mihalcea, R., Apr. 2007. In: Proceedings of NAACL HLT 2007. Rochester, New York, USA, pp. 196–203.
[83] Mihalcea, R., Csomai, A., 2007. Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 233–242.
[84] Mika, P., Ciaramita, M., Zaragoza, H., Atserias, J., Oct. 2008. Learning to tag and tagging to learn: a case study on wikipedia. IEEE Intelligent Systems 23 (5), 26–33.
[85] Milne, D., Medelyan, O., Witten, I. H., 2006. Mining domain-specific thesauri from wikipedia: A case study. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. pp. 442–448.

[86] Milne, D., Witten, I. H., 2008. Learning to link with wikipedia. In: 17th ACM Conference on Information and Knowledge Management, CIKM'08, October 26, 2008 - October 30, 2008. Napa Valley, CA, United states, pp. 509–518.
[87] Milne, D., Witten, I. H., Nichols, D. M., 2007. A knowledge-based search engine powered by wikipedia. In: 16th ACM Conference on Information and Knowledge Management, CIKM 2007, November 6, 2007 - November 9, 2007. Lisboa, Portugal, pp. 445–454.
[88] Muchnik, L., Itzhack, R., Solomon, S., Louzoun, Y., 2007. Self-emergence of knowledge trees: Extraction of the wikipedia hierarchies. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 76 (1).
[89] Murugeshan, M. S., Lakshmi, K., Mukherjee, S., Apr. 2010. A negative category based approach for wikipedia document classification. International Journal of Knowledge Engineering and Data Mining 1, 84–97.
[90] Nielsen, F. r., 2008. Clustering of scientific citations in wikipedia. In: Proceedings of Wikimania 2008.
[91] Nielsen, F. r., Feb. 2012. Wikipedia research and tools: Review and comments.
[92] Okoli, C., 2009. A brief review of studies of wikipedia in peer-reviewed journals. In: Digital Society, 2009. ICDS'09. Third International Conference on. p. 155"160.
[93] Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., Lanamäki, A., Mar. 2012. The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia. SSRN eLibrary.
URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2021326
[94] Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., Lanamäki, A., 2014. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on wikipedia readers and readership. Journal of the Association for Information Science and Technology 65 (12).
[95] Okoli, C., Schabram, K., Oct. 2009. Protocol for a systematic literature review of research on the wikipedia. In: Proceedings of the International Conference on Management of Emergent Digital EcoSystems (MEDES). Lyon, France, p. 73.
[96] Okoli, C., Schabram, K., Sep. 2009. Protocol for a systematic literature review of research on the wikipedia. Sprouts: Working Papers in Information Systems 9 (65).
[97] Okoli, C., Schabram, K., 2010. A guide to conducting a systematic literature review of information systems research. Sprouts: Working Papers on Information Systems 10 (26).
[98] Overell, S., Ruger, S., Mar. 2008. Using co-occurrence models for place name disambiguation. International Journal of Geographical Information Science 22 (3), 265–87.
[99] Overell, S., Sigurbjornsson, B., Van Zwol, R., 2009. Classifying tags using open content resources. In: 2nd ACM International Conference on Web Search and Data Mining, WSDM'09, February 9, 2009 - February 12, 2009. Barcelona, Spain, pp. 64–73.
[100] Pak, A. N., Chung, C.-W., 2010. A wikipedia matching approach to contextual advertising. World Wide Web 13 (3), 251–274.
[101] Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.-M., Vyas, V., 2009. Web-scale distributional similarity and entity set expansion. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2. pp. 938–947.
[102] Pehcevski, J., Thom, J., Vercoustre, A., Naumovski, V., Oct. 2010. Entity ranking in wikipedia: utilising categories, links and topic difficulty prediction. Information Retrieval 13 (5), 568.
[103] Perea-Ortega, J. M., Montejo-Raez, A., Martin-Valdivia, M., Urena-Lopez, L., 2010. Using web sources for improving video categorization, 1–14.
[104] Pöllä, M., Honkela, T., Nov. 2010. Negative selection of written language using character multiset statistics. Journal of Computer Science and Technology 25 (6), 1256–1266.
[105] Ponzetto, S. P., Strube, M., 2006. Exploiting semantic role labeling, WordNet and wikipedia for coreference resolution. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 192 – 199.
[106] Ponzetto, S. P., Strube, M., 2007. Deriving a large scale taxonomy from wikipedia. In: Proceedings: 22nd AAAI Conference on Artificial Intelligence and the 19th Innovative Applications of Artificial Intelligence Conference, July 22-26, 2007. Vol. 2. Vancouver, BC, Canada, pp. 1440–1445.
[107] Ponzetto, S. P., Strube, M., 2007. Knowledge derived from wikipedia for computing semantic relatedness. Journal of Artificial Intelligence Research 30, 181–212.
[108] Potthast, M., Barrn-Cedeo, A., Stein, B., Rosso, P., Jan. 2010. Cross-language plagiarism detection. Language Resources and Evaluation, 1–18.
[109] Quack, T., Leibe, B., Gool, L. V., 2008. World-scale mining of objects and events from community

photo collections. In: Proceedings of the 2008 international conference on Content-based image and video retrieval. pp. 47–56.
[110] Rahurkar, M., Tsai, S.-F., Dagli, C., Huang, T., 2010. Image interpretation using large corpus: Wikipedia. Proceedings of the IEEE 98 (8), 1509–25.
[111] Ray, S. K., Singh, S., Joshi, B., 2010. A semantic approach for question classification using WordNet and wikipedia. Pattern Recognition Letters 31 (13), 1935–1943.
[112] Ruiz-Casado, M., Alfonseca, E., Castells, P., 2007. Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from wikipedia. Data and Knowledge Engineering 61 (3), 484–499.
[113] Schenkel, R., Suchanek, F. M., Kasneci, G., 2007. YAWN: a semantically annotated wikipedia XML corpus. In: 12. Symposium on Database Systems for Business, Technology and the Web of the German Society for Computer Science. Vol. 12. p. 277"291.
[114] Sigurdsson, M., Halling, S. C., 2007. Zeeker: A topic-based search engine. Ph.D. thesis, Informatics and Mathematical Modelling, Technical University of Denmark, Kongens Lyngby, Denmark.
[115] Silva, F., Travencolo, B., Viana, M., da Fontoura Costa, L., 2010. Identifying the borders of mathematical knowledge. Journal of Physics A: Mathematical and Theoretical 43 (32), 325202 (7 pp.).
[116] Simma, A., 2010. Modeling events in time using cascades of poisson processes. Ph.D. thesis, University of California, Berkeley, United States – California.
[117] Stokes, N., Li, Y., Moffat, A., Rong, J., 2008. An empirical study of the effects of NLP components on geographic IR performance. International Journal of Geographical Information Science 22 (3), 247–264.
[118] Stone, B., Dennis, S., Kwantes, P. J., 2010. Comparing methods for single paragraph similarity analysis. Topics in Cognitive Science, no.
[119] Strube, M., Ponzetto, S. P., 2006. WikiRelate! computing semantic relatedness using wikipedia. In: Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence. Menlo Park, Califonia, p. 1419"1424.
[120] Suchanek, F. M., Kasneci, G., Weikum, G., 2007. YAGO: a core of semantic knowledge unifying WordNet and wikipedia. In: Proceedings of the 16th international conference on World Wide Web. pp. 697 – 706.
[121] Syed, Z., 2010. Wikitology: A novel hybrid knowledge base derived from wikipedia. Ph.D. thesis, University of Maryland, Baltimore County, United States – Maryland.
[122] Tan, B., Peng, F., 2008. Unsupervised query segmentation using generative language models and wikipedia. In: Proceeding of the 17th international conference on World Wide Web. pp. 347–356.
[123] Theobald, M., Bast, H., Majumdar, D., Schenkel, R., Weikum, G., 2008. TopX: efficient and versatile top-k query processing for semistructured data. The VLDB Journal " The International Journal on Very Large Data Bases 17 (1), 81 – 115.
[124] Turdakov, D., Kuznetsov, S., 2010. Automatic word sense disambiguation based on document networks. Programming and Computer Software 36 (1), 11–18.
[125] Turdakov, D., Velikhov, P., 2008. Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation. In: Spring Young Researcher's Colloquium On Database and Information Systems. St.-Petersburg, Russia.
[126] Vechtomova, O., 2010. Facet-based opinion retrieval from blogs. Information Processing and Management 46 (1), 71–88.
[127] Řehůřek, R., Jun. 2010. Fast and faster: A comparison of two streamed matrix decomposition algorithms. In: NIPS 2010 Workshop on Low-rank Methods for Large-scale Machine Learning. Vancouver, Canada.
[128] Wang, P., Domeniconi, C., 2008. Building semantic kernels for text classification using wikipedia. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 713–721.
[129] Wang, P., Hu, J., Zeng, H.-J., Chen, Z., 2009. Using wikipedia knowledge to improve text classification. Knowledge and Information Systems 19 (3), 265–281.
[130] Wang, Y.-C., Tsai, R. T.-H., Hsu, W.-L., Mar. 2009. Web-based pattern learning for named entity translation in Korean"Chinese cross-language information retrieval. Expert Systems with Applications 36 (2, Part 2), 3990–3995.
[131] Weiss, S., Urso, P., Molli, P., 2010. Logoot-undo: Distributed collaborative editing system on P2P networks. IEEE Transactions on Parallel and Distributed Systems 21 (8), 1162–1174.
[132] Weld, D. S., Wu, F., Adar, E., Amershi, S., Fogarty, J., Hoffmann, R., Patel, K., Skinner, M., 2008. Intelligence in wikipedia. In: Proceedings of the 23rd national conference on Artificial intelligence

- Volume 3. pp. 1609–1614.
[133] Wong, W., Liu, W., Bennamoun, M., 2007. Tree-traversing ant algorithm for term clustering based on featureless similarities. Data Mining and Knowledge Discovery 15 (3), 349–381.
[134] Wu, F., Hoffmann, R., Weld, D. S., 2008. Information extraction from wikipedia: Moving down the long tail. In: 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008, August 24-27, 2008. Las Vegas, NV, United states, pp. 731–739.
[135] Wu, F., Weld, D. S., 2007. Autonomously semantifying wikipedia. In: 16th ACM Conference on Information and Knowledge Management, CIKM 2007, November 6, 2007 - November 9, 2007. Lisboa, Portugal, pp. 41–50.
[136] Xiang, E. W., Cao, B., Hu, D. H., Yang, Q., 2010. Bridging domains using world wide knowledge for transfer learning. IEEE Transactions on Knowledge and Data Engineering 22 (6), 770–783.
[137] Yu, J., Thom, J. A., Tam, A., 2009. Requirements-oriented methodology for evaluating ontologies. Information Systems 34 (8), 686–711.
[138] Zaragoza, H., Rode, H., Mika, P., Atserias, J., Ciaramita, M., Attardi, G., 2007. Ranking very many typed entities on wikipedia. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. Lisbon, Portugal, pp. 1015–1018.
[139] Zesch, T., Gurevych, I., Jan. 2010. Wisdom of crowds versus wisdom of linguists - measuring the semantic relatedness of words. Natural Language Engineering 16 (1), 25–59.
[140] Zesch, T., Müller, C., Gurevych, I., 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation (LREC).
[141] Zesch, T., Müller, C., Gurevych, I., 2008. Using wiktionary for computing semantic relatedness. In: Proceedings of the 23rd national conference on Artificial intelligence - Volume 2. pp. 861–866.
[142] Zhang, X., 2009. Exploiting external/domain knowledge to enhance traditional text mining using graph-based methods. Ph.D. thesis, Drexel University, United States – Pennsylvania.
[143] Zhirov, A. O., Zhirov, O. V., Shepelyansky, D. L., Oct. 2010. Two-dimensional ranking of wikipedia articles. The European Physical Journal B - Condensed Matter and Complex Systems, 1–9.
[144] Zhou, A., Zhang, R., Qian, W., Vu, Q. H., Hu, T., 2008. Adaptive indexing for content-based search in P2P systems. Data and Knowledge Engineering 67 (3), 381–398.