Information Processing and Management 000 (2016) 1–25

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

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[e]

[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

ARTICLE INFO

Article history:
Received 30 November 2014
Revised 19 July 2016
Accepted 28 July 2016
Available online xxx

Keywords:
Information retrieval
Information extraction
Natural language processing
Ontologies
Wikipedia
Literature review

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.

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 (Mesgari, Okoli, Mehdi, Nielsen, & Lanamäki, 2015; Okoli, Mehdi, Mesgari, Nielsen, & Lanamäki, 2012, 2014). 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, Milne, Legg, and Witten (2009) 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 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

* Corresponding author.
E-mail addresses: mo_mehdi@encs.concordia.ca, mohamad_mehdi@hotmail.com (M. Mehdi), Chitu.Okoli@concordia.ca (C. Okoli).

http://dx.doi.org/10.1016/j.ipm.2016.07.003
0306-4573/

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 (Okoli et al., 2012). 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 (Mesgari et al., 2015), and explored its readers and readership behaviors (Okoli, Mehdi, Mesgari, Nielsen, & Lanamäki, 2014). One of the most thorough literature reviews of Wikipedia is that of Medelyan et al. (2009), 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*" Medelyan et al. (2009, 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 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. (2009) 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 (2010). 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 (Okoli & Schabram, 2009a). 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 (Okoli et al., 2012). The selected studies are categorized according to their

Please cite this article as: M. Mehdi et al., Excavating the mother lode of human-generated text: A systematic review of research that uses the wikipedia corpus, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.07.003