Information Processing and Management 50 (2014) 284–296

Contents lists available at ScienceDirect

Information Processing and Management

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

Preferences in Wikipedia abstracts: Empirical findings and implications for automatic entity summarization

Danyun Xu, Gong Cheng *, Yuzhong Qu

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, PR China

ARTICLE INFO

Article history:
Received 10 September 2013
Received in revised form 13 November 2013
Accepted 23 December 2013
Available online 19 January 2014

Keywords:
DBpedia
Entity summarization
Feature selection
Property ranking
Wikipedia

ABSTRACT

The volume of entity-centric structured data grows rapidly on the Web. The description of an entity, composed of property-value pairs (a.k.a. features), has become very large in many applications. To avoid information overload, efforts have been made to automatically select a limited number of features to be shown to the user based on certain criteria, which is called automatic entity summarization. However, to the best of our knowledge, there is a lack of extensive studies on how humans rank and select features in practice, which can provide empirical support and inspire future research. In this article, we present a large-scale statistical analysis of the descriptions of entities provided by DBpedia and the abstracts of their corresponding Wikipedia articles, to empirically study, along several different dimensions, which kinds of features are preferable when humans summarize. Implications for automatic entity summarization are drawn from the findings.

1. Introduction

Entity-centric structured data such as Google's Knowledge Graph, Facebook's Open Graph, and W3C's RDF and Linked Data, has become an important component of the Web. It describes the attributes of and the relationships between entities. The description of an entity, or an *entity description* for short, is composed of property-value pairs, a.k.a. *features* (Cheng, Tran, & Qu, 2011), as illustrated by the right-hand side of Fig. 1. The value of a property could be a data value (e.g. an integer), another entity, or a class (usually being the type of the entity). The volume of such data increases rapidly on the Web. Entities are associated with more and more features. For instance, the RDF description of Sydney provided by Freebase[1] contains several hundred features. When showing such an entity description to the user, to avoid information overload, practical applications like Knowledge Graph on Google's search results pages present not all but only a limited number (e.g. top-k) of features, called a summary of this entity description, or an *entity summary* for short. Then, the problem arises as to which features are best for constituting an entity summary to be used in a particular application (or, how to rank features), and the term *entity summarization* was coined to describe this problem (Cheng et al., 2011).

This emerging problem has received attention from researchers in the areas of information retrieval (Zhang, Zhang, & Chen, 2012), database (Fakas, 2011), and Semantic Web (Cheng et al., 2011). To develop an effective approach to entity summarization, a key issue to consider is how humans summarize entity descriptions by ourselves. However, to the best of our knowledge, there is a lack of extensive empirical studies on this topic, which motivates our work. In this article, we aim to explore, via a large-scale empirical analysis, which kinds of features are preferable when humans summarize entity descriptions for generic use. Implications for developing approaches to automatic entity summarization will be drawn from the findings.

* Corresponding author. Tel./fax: +86 (0)25 89680923.
E-mail addresses: dyxu@smail.nju.edu.cn (D. Xu), gcheng@nju.edu.cn (G. Cheng), yzqu@nju.edu.cn (Y. Qu).

[1] http://www.freebase.com/m/06y57. Last accessed: 08/31/2013.

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.12.001

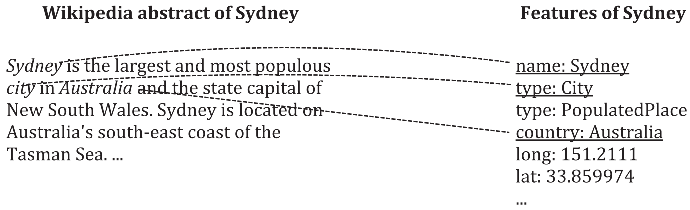

Fig. 1. Identification of the features (underlined) mentioned in the corresponding Wikipedia abstract.

To achieve this, instead of inviting a few human experts to summarize entity descriptions, which can hardly be extended to a large scale, we intend to analyze millions of entity descriptions and the summaries thereof given by a large and representative population. Therefore, we choose DBpedia (Bizer et al., 2009), a well-known dataset at the center of Linked Data. It treats the topic of each Wikipedia article as an entity, and extracts its structured description from the article. The results are represented as *RDF triples* (Klyne & Carroll, 2004), each of which comprises an entity, a property, and a value, or in other words, a feature of an entity, as illustrated by the right-hand side of Fig. 1. Thanks to the encyclopedic topics and massive amount of information offered by Wikipedia, DBpedia has collected tens of millions of RDF triples describing several million entities. To obtain general-purpose summaries of these entity descriptions, we exploit the first section of each Wikipedia article (i.e. the text prior to the table of contents), called a *Wikipedia abstract*, which provides an abstract of the article and thus is regarded as a generic textual summary of the corresponding entity description, as illustrated by the left-hand side of Fig. 1. Then, we identify the features in each entity description that are mentioned in its corresponding Wikipedia abstract, as illustrated by Fig. 1, by using an automatic approach. The results constitute an entity summary, with which we can analyze and reveal humans' preferences in choosing features. In particular, these preferences belong to not a small population but the large Wikipedia community, thereby making our findings more generalizable.

Our major contributions are:

- a large-scale multi-dimensional statistical analysis of humans' preferences in selecting features into an entity summary, and
- a number of implications drawn for automatic entity summarization.

To be specific, we will investigate the length of entity summaries, analyze the priorities of, diversity of, and correlation between properties in summarization, and explore the preferences in choosing property values. Based on the empirical findings, several heuristics are recommended to be considered in future research on entity summarization.

The remainder of this article is organized as follows. Section 2 reviews the literature. Section 3 describes the dataset to use. Section 4 introduces and evaluates several strategies for identifying which features are mentioned in a Wikipedia abstract. Section 5 presents a multi-dimensional statistical analysis of the features mentioned in Wikipedia abstracts to explore humans' preferences, from which implications for automatic entity summarization are drawn in Section 6. Finally, Section 7 concludes the article with a discussion of future work.

2. Literature review

2.1. Entity summarization

Entity summarization has proven to be important to entity search engines (Bai, Delbru, & Tummarello, 2008; Cheng & Qu, 2009), where *query-biased entity summaries*, consisting of features that contain query keywords, are created to be shown on search engine results pages. Zhang et al. (2012) used a learning-based approach to rank features according to their relevance to the query. Google (2012), to summarize an entity description in Knowledge Graph, may have utilized query logs to find the properties that have been asked more often in Google Search.

Recent interests mainly focused on generating a summary that can best characterize the underlying entity *for generic use*. RELIN (Cheng et al., 2011) employs a random surfer model to rank features according to their informativeness and relatedness. The measurement of informativeness is based on the self-information of features, and the measurement of relatedness is based on the co-occurrence of properties and of values on the Web. Thalhammer, Toma, Roa-Valverde, and Fensel (2012) considered a feature of an entity important if it is shared with the entity's nearest neighbors. The distance between entities is defined based on their rates given by human users in certain applications. DIVERSUM (Sydow, Pikuła, & Schenkel, 2013) extends the notion of entity summary to be an arbitrary connected subgraph surrounding the entity in graph-structured data. A summary is constructed in a greedy manner by successively adding the edge that has both a short distance to the entity and a property frequently used in the data. In particular, features sharing a common property are not allowed to appear in a summary together, to improve the diversity.

The database community tackle a similar problem when providing keyword search in relational databases. As Fakas (2011) discussed, a single query-relevant tuple returned by keyword search does not comprise a complete result; additional