A Method for Predicting Wikipedia Editors’ Editing Interest Based on a Factor Graph Model

Haisu Zhang, Academy of National Defense Information, Wuhan, China
Sheng Zhang, Academy of National Defense Information, Wuhan, China
Zhaolin Wu, Academy of National Defense Information, Wuhan, China
Liwei Huang, Beijing Institute of Remote Sensing, Beijing, China
Yutao Ma, Wuhan University, Wuhan, China & WISET Automation Co., Ltd., Wuhan, China

ABSTRACT

Recruiting or recommending appropriate potential Wikipedia editors to edit a specific Wikipedia entry (or article) can play an important role in improving the quality and credibility of Wikipedia. According to empirical observations based on a small-scale dataset collected from Wikipedia, this paper proposes an Interest Prediction Factor Graph (IPFG) model, which is characterized by editor’s social properties, hyperlinks between Wikipedia entries, the categories of an entry and other important features, to predict an editor’s editing interest in types of Wikipedia entries. Furthermore, the paper suggests a parameter learning algorithm based on the gradient descent algorithm and the Loopy Sum-Product algorithm for factor graphs. An experiment on a Wikipedia dataset (with different frequencies of data collection) shows that the average prediction accuracy (F1 score) of the IPFG model for data collected quarterly could be up to 0.875, which is approximately 0.49 higher than that of a collaborative filtering approach. In addition, the paper analyzes how incomplete social properties and editing bursts affect the prediction accuracy of the IPFG model. The authors’ results can provide insight into effective Wikipedia article tossing and can improve the quality of special entries that belong to specific categories by means of collective collaboration.

KEYWORDS

Factor Graph, Interest Prediction, Probabilistic Graphical Model, Social Network Mining, Wikipedia

1. INTRODUCTION

Wikipedia has become the largest free encyclopedia on the Internet, and it is built by numerous volunteer editors. As of August 2014, there are more than 4.59 million English entries (also called articles) on Wikipedia, all of which are cooperatively edited by volunteers rather than by professional domain experts, without too many mandatory editing constraints. Although Wikipedia has garnered some recognition for high-quality entries—e.g., in 2005, an article in the journal Nature declared that Wikipedia was “nearly as accurate as Britannica” on many scientific topics (Giles, 2005)—many people express concern about the quality and credibility of randomly selected Wikipedia entries (especially for the byways of some subjects).
As we know, for an open Web 2.0 system such as Wikipedia, the quality of entries depends largely on the wisdom of crowds (Kittur and Kraut 2008) because large numbers of editors contribute to the accuracy, completeness and clarity of those entries. As a result of sparse hyperlinks to other frequently used Wikipedia entries, inadequate text content, rare domain knowledge and a number of other reasons, there are some Wikipedia entries that receive very little attention from editors. Thus, it is sometimes difficult for the right editor(s) to find entries that need further editing, likely leading to low-quality contributions to the development of Wikipedia (Wilkinson and Huberman 2007). As far as we are concerned, predicting possible volunteer editors for a specific Wikipedia entry, or recommending the appropriate Wikipedia entries to a given editor according to his/her editing interest, could provide a more effective mode of collaborative editing, improving the quality and credibility of the above-mentioned special types of Wikipedia entries.

In view of historical editing data, including editors and their corresponding Wikipedia entries, there are several approaches to predicting an editor’s editing interest, such as collaborative recommendations, classification algorithms, and probabilistic graphical models. A probabilistic graphical model is a probabilistic model where the conditional dependencies between the random variables are specified via a graph. The framework of probabilistic graph models, which provides algorithms for finding and analyzing structure in complex distributions to describe them simply and extract unstructured information, makes it possible for them to be built and used effectively (Koller and Friedman 2009). Because of the advantages over other methods (e.g., inference and learning are treated together, and supervised and unsupervised learning are merged seamlessly), probabilistic graph models have recently been widely utilized in many machine learning algorithms, including Bayesian Networks, Markov Networks, and Template Models.

To recommend specific Wikipedia entries that receive little attention to potentially appropriate editors, the main goal of this paper is to build a factor graph for predicting editors’ editing interest (also called the Interest Prediction Factor Graph (IPFG) model), which is a specific presentation of probabilistic graph models. Moreover, we examined and validated the feasibility and superiority of our method with empirical experiments on a dataset collected from Wikipedia. In general, the primary technical contributions of the paper are described as follows:

1. Unlike traditional item-based or user-based recommendation methods, we proposed a novel learning model based on factor graphs to predict editors’ editing interest. The IPFG model takes editors’ social properties, hyperlinks between Wikipedia entries, the categories of an entry, and other important features into account, and the parameters of the model are learned by the gradient descent algorithm and the Loopy Sum-Product algorithm. The experimental results show that our model does outperform a collaborative filtering recommendation algorithm on prediction accuracy in terms of the F-measure;

2. On the basis of the proposed prediction model, on one hand, we analyzed the impact of a different number of editors’ social properties on the prediction accuracy of our model and found that hobbies and profession have the greatest impacts on prediction accuracy; on the other hand, we also analyzed how different time steps (i.e., data collection frequencies) of the history of editors’ cooperative behavior, especially for editing bursts on Wikipedia, affect the prediction accuracy of our model and found that the models for data collected quarterly and weekly achieve the highest and lowest prediction accuracy, respectively.

The rest of this paper is organized as follows: Section 2 reports our empirical observations of behavior homophily on Wikipedia according to a small-scale dataset, Section 3 presents a formal description of the problem of predicting editors’ editing interest, Section 4 proposes the framework of the IPFG model, Section 5 introduces the learning algorithm for the parameters of our model, Section 6 discusses the experimental results on a dataset gathered from Wikipedia with five different
A Bounded Model Checking Approach for the Verification of Web Services Composition

Model-Driven Semantic Web Services