Data & Knowledge Engineering 92 (2014) 39–59

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Editorial

A generic framework and methodology for extracting semantics from co-occurrences

Aditya Ramana Rachakonda *, Srinath Srinivasa, Sumant Kulkarni, M.S. Srinivasan [1]

Open Systems Lab, IIIT Bangalore, India

ARTICLE INFO

Article history:
Received 15 June 2013
Received in revised form 24 May 2014
Accepted 17 June 2014
Available online 27 June 2014

Keywords:
Cognitive models
Co-occurrence
Data mining
Text mining

ABSTRACT

Extracting semantic associations from text corpora is an important problem with several applications. It is well understood that semantic associations from text can be discerned by observing patterns of co-occurrences of terms. However, much of the work in this direction has been piecemeal, addressing specific kinds of semantic associations. In this work, we propose a generic framework, using which several kinds of semantic associations can be mined. The framework comprises a co-occurrence graph of terms, along with a set of graph operators. A methodology for using this framework is also proposed, where the properties of a given semantic association can be hypothesized and tested over the framework. To show the generic nature of the proposed model, four different semantic associations are mined over a corpus comprising of Wikipedia articles. The design of the proposed framework is inspired from cognitive science — specifically the interplay between semantic and episodic memory in humans.

1. Introduction

Extracting latent semantics from human generated text corpora, like a collection of news articles, email and blog post and so on is an important problem with several application areas. A practical approach to acquire such latent semantics is to observe patterns of term distribution. It is well known that term distributions in human generated text are not independent of each other and sets of semantically related terms tend to occur together [1]. In linguistics, this correlation is termed as the *distributional hypothesis* [2]. Because of this, observing occurrence and co-occurrence patterns of terms has become the primary mode of gathering latent semantics from unstructured text. Some examples include [3–6].

However, to the best of our knowledge, much of the research efforts in this direction, has been piecemeal solutions focused on identifying specific kinds of semantic associations.

In this work, we try to build a theory around the question of how semantic associations can be discerned from co-occurrences and propose a generic framework for using co-occurrence patterns to extract different kinds of semantic associations.

In the proposed framework, each document in a corpus is treated as an "episode" and a raw form of semantic knowledge is represented as a weighted graph showing co-occurrences of terms across the corpus. A set of graph operations called *primitives* is also proposed using which, several kinds of co-occurrence patterns can be discerned. Finally, a methodology is proposed, where a hypothesis about a semantic association can be tested on the co-occurrence graph using the primitives. The validity of the hypothesis is established by testing the outcome of an algorithm with human subjects, as well as with comparisons with related approaches.

* Corresponding author at: Q801, Ajmera Infinity, Electronics City, Bangalore, India 560100.
E-mail addresses: aditya.rachakonda@gmail.com (A.R. Rachakonda), sri@iiitb.ac.in (S. Srinivasa), sumant@iiitb.ac.in (S. Kulkarni), srinivasan.ms@iiitb.org (M.S. Srinivasan).

[1] Affiliated with IBM India Software Labs and with IIIT Bangalore as a part-time research scholar.

http://dx.doi.org/10.1016/j.datak.2014.06.002
0169-023X/© 2014 Elsevier B.V. All rights reserved.

Unlike traditional natural language processing, in this work, we do not rely on rule-based text processing techniques like part-of-speech tagging, dependency trees and so on. Instead we assume such techniques to be unavailable and use free-text along with statistical keyword detection algorithms to measure co-occurrences. The underlying data structures and primitives are not dependent upon the semantic associations proposed and are constant across all the algorithms.

While the techniques used in the specific algorithms (like random walks, clustering, K–L divergence, etc.) are not novel by themselves, the focus of this work is on the generic nature of the proposed framework. A single data structure along with a small set of graph primitives, can be used to mine different kinds of semantic associations by following a methodology. This way, there is a basis for explaining the rationale behind a semantic association that is mined from the corpus.

The rest of the paper is organized as follows. Section 2 briefly describes related literature in mining latent semantics from text. Section 3 proposes the 3-layer framework for mining semantic associations. Four different latent semantic associations based on this model are then demonstrated in Section 4 and the concluding remarks are noted in Section 5.

2. Related literature

In this section, we survey literature in mining latent semantics. The objective here is to provide a realistic backdrop for this work, rather than a comprehensive survey. For the latter, the interested reader may like to refer to [7–11].

2.1. Co-occurrence graph mining

Co-occurrences represented as a graph are one of the fundamental structures using which the distributional hypothesis can be applied to semantics mining. The way co-occurrence graphs are constructed, are usually algorithm specific, as different kinds of co-occurrences are thought to capture different aspects of meaning.

In the context of word-sense disambiguation, Widdows et al. use lists of nouns from a part-of-speech (POS) tagged corpus to construct a co-occurrence graph [12,13]. They then use a graph clustering algorithm to identify significant clusters of words and thereby distinguishing between word senses. In topic mining, Mihalcea and Tarau [14] propose a document-level co-occurrence graph of terms (nouns and adjectives) inside a document and compute a random-walk centrality of the nodes to identify the terms representing the topic of the document. They also show that the same technique can be used on sentences in a document to identify the sentence summarizing the document best.

There are also several algorithms which use a co-occurrence graph consisting of explicitly heterogeneous nodes connected in a k-partite arrangement. A *noun–adjective* bipartite co-occurrence graph can be used to determine the sentiments associated with the nouns [15]. A bipartite random-walk on such a graph is used to identify important opinions (adjectives) and important product features (nouns) in a given product corpus. Similarly, a *noun–verb* bipartite co-occurrence graph can be used to mine semantics especially in speech disambiguation [16].

In our work, we propose a set of semantics mining algorithms on a *single* co-occurrence graph of terms built over a large textual corpus.

2.2. Dimensionality reduction

An alternate approach to mining latent semantics has focused on discovering non-trivial latent co-occurrences by means of dimensionality reduction as in Latent Semantic Analysis (LSA) [17]. LSA uses singular value decomposition on a term–document matrix, then collapses the vector space by eliminating all but the top k dimensions and hence document vectors which were far off in the original space come closer in the new space. Such a recomputed space establishes extraneous associations between documents and terms beyond what was originally captured.

Dimension reduction techniques have also been applied to co-occurrence graph mining in Hyperspace Analogue to Language (HAL) [18] and Correlated Occurrence Analogue to Lexical Semantics (COALS) [3]. Both the algorithms work on a term–term matrix composed of co-occurrence vectors instead of document vectors for mining semantic associations.

Despite their impressive results, LSA and its variants do not have sound mathematical underpinnings for the extracted semantics. This meant that, while terms could be semantically related by collapsing dimensions, LSA is not able to assign a label to such associations. In addition, LSA computations are global in nature involving the entire corpus, making it difficult for incremental changes in computing semantic relatedness. Several research efforts have tried to extend LSA in different directions as well as explore newer models for capturing latent semantics.

2.3. Generative models

Generative models in semantics started as a mathematically sound extension to LSA. Here, documents in the corpus are considered to be generated by a mixture of one or more random processes.

Hofmann proposed pLSI [19], a probabilistic approach for topical mixture model. Here, a document is modeled as comprising a set of topics, where each topic generates terms with a given probability distribution.

Latent Dirichlet Allocation (LDA) [20] is an extension of PLSI, where a document is modeled as generated using a mixture of k (finite) hypothetical topics and a topic is a probability distribution over all the terms observed in the corpus, based on a Dirichlet