Natural Language Engineering **21** (5): 725–742. © Cambridge University Press 2015
doi:10.1017/S135132491500008X

Clique-based semantic kernel with application to semantic relatedness

A. H. JADIDINEJAD
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
e-mail: a.jadidi@srbiau.ac.ir

F. MAHMOUDI
Computer and IT Engineering Faculty, Islamic Azad University, Qazvin Branch, Qazvin, Iran

M. R. MEYBODI
Computer Engineering and Information Technology Department, Amirkabir University of Technology, Tehran, Iran

(*Received 20 May 2014; revised 24 January 2015; accepted 26 January 2015; first published online 14 April 2015*)

Abstract

The emergence of knowledge repositories in a variety of domains provides a valuable opportunity for semantic interpretation of high dimensional datasets. Previous researches investigate the use of concept instead of word as a core semantic feature for incorporating semantic knowledge from an ontology into the representation model of documents. On the other hand, in machine learning and information retrieval, data objects are represented as a flat feature vector. The inconsistency between the structural nature of the knowledge repositories and the flat representation of features in machine learning leads researchers to neglect the structure of the knowledge base and leverage concepts as isolated semantic features, which is known as bag-of-concepts. Although, using concepts has some advantages over words, by neglecting the relation between concepts, the problem of vocabulary mismatch remains in force. In this paper, a novel semantic kernel is proposed which is capable of incorporating the relatedness between conceptual features. This kernel leverages clique theory to map data objects to a novel feature space wherein complex data objects will be comparable. The proposed kernel is relevant to all applications which have a prior knowledge about the relatedness between features. We concentrate on representing text documents and words using Wikipedia and WordNet, respectively. The experimental results over a set of benchmark datasets have revealed that the proposed kernel significantly improves the representation of both words and texts in the application of semantic relatedness.

1 Introduction

In many applications of machine learning, data objects are represented via a large number of features or dimensions; so-called high dimensional data in the literature (Kriegel, Peer and Zimek 2009; Assent 2012). Most of these features are typically related. For example, when using words as features to represent a

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

textual document, it is inevitable that different documents leverage different words to mention a same concept. Previous researches (Kriegel *et al.* 2009; Assent 2012) have revealed that the concept of distance becomes less precise as the number of related features (dimensions) grows, i.e. in high dimensions all instances look alike. On the other hand, the assessment of relatedness using distances plays an important role in different applications of machine learning such as clustering or classification. Consequently, recent researches have focused on developing techniques for representing high dimensional data (Assent 2012).

Traditional bag-of-words is still a well-known representation model (Turney and Pantel 2010). In this model, each word/document is represented as a high dimensional vector with a dimension for each term of the dictionary. This model is incapable of representing multi-words, synonymy and polysemy (Wang and Domeniconi 2008). In most applications, there is a prior knowledge about the relatedness of semantic features; or the relatedness of the features can be inferred. For example, by employing current repositories such as WordNet (Finlayson 2014), the relatedness of the words can be calculated (Zhang, Gentile and Ciravegna 2013).

On the other hand, the emergence of knowledge repositories such as Wikipedia (Medelyan *et al.* 2009), DBpedia (Bizer *et al.* 2009) and BabelNet (Navigli and Ponzetto 2012) provides a valuable opportunity for semantic interpretation. Concept-based information retrieval leverages these ontologies and goes beyond the word level and works with concepts instead (Huang *et al.* 2012; Jadidinejad and Mahmoudi 2014). According to the leveraged ontology, the nature of concepts will be Wikipedia articles, WordNet synsets and etc. Using concepts instead of words are essential for solving knowledge-intensive tasks such as semantic relatedness. Measuring semantic relatedness is an important task in natural language processing. It is often a pre-processing step for many applications such as clustering, classification and information retrieval (Zhang *et al.* 2013).

In machine learning, data objects are represented as flat feature vectors. This inconsistency between the structural nature of the knowledge repositories and the flat representation of features in machine learning and information retrieval leads researchers to neglect the structure of the knowledge base and leverage concepts as isolated features, which is known as bag-of-concepts (Gabrilovich and Markovitch 2009; Huang *et al.* 2012; Jadidinejad and Mahmoudi 2014). Although using concepts instead of words overcomes the limitations of the traditional bag-of-words model, the challenge of related features remains intact. On the other hand, the main important attribute of a concept is its relationship to other concepts. This possibility is completely neglected in the bag-of-concept representation model.

For example, consider two documents in Table 1. The most important extracted Wikipedia concepts are shown in bold. According to the provided human judgments, these documents are highly related to each other ($r = 4.33/5$). Unfortunately, there is no distinguished exact match between the extracted concepts or words. So, bag-of-concepts or bag-of-words representation models are incapable to detect the amount of semantic relatedness between them. On the other hand, both documents are about global environment and the role of human in relation to it. In this paper, we propose a novel representation model in the form of semantic kernel which is

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

Table 1. *Two example documents from Lee data set. Extracted Wikipedia concepts have been shown in* ***bold****. Although these documents are highly relevant (*4.33/5*), there is no exact match between the extracted concepts*

Document #40	Document #43
The real level of world inequality and **environmental degradation** may be far worse than official estimates, according to a leaked document prepared for the world's richest countries and seen by **the Guardian**. It includes new estimates that the world lost almost 10% of its **forests** in the past 10 years; that **carbon dioxide emissions** leading to **global warming** are expected to rise by 33% in rich countries and 100% in the rest of the world in the next 18 years; and that more than 30% more fresh water will be needed by 2020.	**Pope John Paul II** urged delegates at a major **U.N.** summit on **sustainable growth** on Sunday to pursue development that protects the **environment** and **social justice**. In comments to tourists and the faithful at his summer residence southeast of **Rome**, the pope said **God** had put humans on **Earth** to be his administrators of the land, 'to cultivate it and take care of it.' 'In a world ever more interdependent, **peace**, **justice**, and the safekeeping of creation cannot but be the fruit of a joint commitment of all in pursuing the **common good**,' John Paul said.

capable of incorporating the relatedness between concepts. The proposed semantic kernel leverages *cliques-of-concepts* instead of explicit bag-of-concepts to represent a textual document[1]. For example, suppose that ('Global warming', 'Earth Summit', 'Natural environment') is a 3-vertex clique in the background knowledge base. This clique provides a cluster of highly related concepts according to the structure of the background knowledge base. The occurrence of 'Global warming' in the first document (document #40) and the occurrence of 'Natural environment' in the second document (document #43), provides a stimulus of semantic relatedness between these two documents. When the number of relatedness stimuli increases, it is more probable that two documents are related. Therefore, using cliques of concepts instead of isolated concepts to represent textual documents allows semantically related documents to be matched. So, it is competent to map data objects to a novel feature space in which the neighborhood among data objects becomes more meaningful. The proposed kernel is relevant to all applications which have a prior knowledge about the relatedness between features; or the relatedness between features can be inferred in any way. The proposed approach is validated in computing semantic relatedness between words and texts. The experimental results over a set of benchmark datasets have revealed that the proposed kernel significantly improves the representation of both words and texts in the application of semantic relatedness.

2 Literature review

Vector space family of kernel methods (Shawe-Taylor and Cristianini 2004), is still the most commonly used representation model in machine learning and information

[1] A clique in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge. It represents a cluster of highly related concepts.

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

retrieval. The original vector space kernel neglects the semantic content. One of the most important issues is to improve the vector space representation to ensure that it is able to support the semantic content of the words (Baroni and Lenci 2010). Kernel methods play an important role in incorporating *a priori* knowledge about both syntactic and semantic structure of terms (Bloehdorn and Moschitti 2007; Mehdad, Moschitti and Zanzotto 2010). For the case of syntactic structure, tree kernels (Croce *et al.* 2011) have been proposed as a framework for exploiting the parse trees of the input texts. For the case of lexical semantics, semantic kernels exploit background information from semantic networks such as WordNet (Basili, Cammisa and Moschitti 2006).

The original vector space considers only the simplest case of linear transformations of the type $\tilde{\phi}(o) = \phi(o) K_m$, where K_m is any $m \times l$ matrix which maps each object from the original m-dimensional space into a new l-dimensional feature space. Different choices of the matrix K_m lead to different variants of vector space semantic kernels. Generalized vector space kernel (Shawe-Taylor and Cristianini 2004) aims at capturing term-term correlations by looking at co-occurrence information. Two terms are considered semantically related if they co-occur in the same document. On the other hand, a document is represented by the embedding $\tilde{\phi}(o) = \phi(o) \acute{D}$, where D is the document-term matrix. This means that two documents can be seen similar even if they do not share any terms.

Latent semantic kernel, aka latent semantic analysis or LSA (Cristianini, Shawe-Taylor and Lodhi 2002) provides a very effective representation of high dimensional data that is able to capture semantic information through the use of co-occurrence information. This approach leverages the singular value decomposition of the matrix $\acute{D} = U \Sigma \acute{V}$, where Σ is a diagonal matrix of the same dimensions as D. U and V are unitary matrices whose columns are the eigenvectors of $\acute{D}D$ and $D\acute{D}$ respectively. Latent semantic kernel represents each object in a new k-dimensional feature vector space $\tilde{\phi}(o) = \phi(o) U_k$, where U_k is the first k columns of the matrix U.

On the other hand, researches have been done to exploit knowledge repositories for concept-based document representation models (Jadidinejad and Mahmoudi 2014). In particular, Wikipedia (Medelyan *et al.* 2009) has been widely used. Gabrilovich *et al.* (2009) proposed a feature generation technique for text documents using Wikipedia, so-called Explicit Semantic Analysis (ESA). In contrast with generalized vector space, ESA leverages an external document space (Anderka and Stein 2009). In this approach, each word is represented as a vector in the high dimensional space defined by Wikipedia articles. Previous research (Liberman and Markovitch 2009) revealed that the main drawback of ESA is its flat representation where each Wikipedia concept is considered as an isolated feature.

Basili *et al.* (2006) proposed a semantic kernel to use the WordNet hierarchy in learning algorithms for document classification. Since the proposed classifier leverages more effective document similarities based on prior knowledge, it is applicable in situations when few training examples are available. Wang and Domeniconi (2008) and Wang *et al.* (2009) overcame the shortfalls of bag-of-words representation model by embedding background knowledge derived from Wikipedia into a semantic kernel, which is then used to enrich the representation of

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

documents. They used a linear combination of three measures of article similarity in Wikipedia in order to extract the proximity matrix of articles. Finally, each document is represented as $\tilde{\phi}(o) = \phi(o) K_m$, where K_m is a semantic matrix which contains both corpus terms and Wikipedia articles.

Fodeh, Punch and Tan (2011) proposed the use of semantic features for incorporating semantic knowledge from an ontology (especially WordNet) into document clustering. They leveraged ontological semantic features to reduce the number of dimensions in the original feature space. Huang *et al.* (2012) proposed a supervised document similarity measure that leverages 17 lexical and semantic features (contains the traditional bag-of-words and the recent bag-of-concepts) to assess document similarity, and learn from human judgments how to combine them by using machine-learning techniques.

Bloehdorn *et al.* (2006) proposed semantic smoothing kernel based on super concept representation. The aim of their work is to embed the knowledge of the topological relations in their super concept expansion into the semantic networks in kernel functions. Semantic smoothing kernel can be defined as a kernel for two data items $x,\ z$ is given by $k(x,\ z) = \acute{x}Qz$ where Q is a square symmetric matrix whose entries represent the semantic proximity between the dimensions of the input space.

3 Clique-based semantic kernel

One of the appeals of using clique-based semantic kernel for pattern analysis is that it can be defined on non-isolated semantic features, make it possible to extend the usefulness of the technique to many different types of applications. All that is required is some measures of similarity that capture the association between two semantic features. The proposed method uses cliques in the feature similarity graph to create more complex relations between objects.

Definition 1: The *object-feature matrix* (so-called *data matrix* in the literature (Shawe-Taylor and Cristianini 2004)) is the matrix $D_{n\times m}$ where rows are indexed by the objects of the collection $o = \{o_1,\ o_2,\ \ldots,\ o_n\}$ and the columns correspond to semantic features $f = \{f_1,\ f_2,\ \ldots,\ f_m\}$. This is a binary matrix which (i, j)th entry gives the existence of feature f_j in object o_i.

Using data matrix (D), an object can be represented as a vector in a high dimensional space in which each dimension is associated with one feature from the feature space (so-called vector space in the literature) (Shawe-Taylor and Cristianini 2004). Equation 1 shows the mapping of a data object o to the corresponding feature space $\phi(o)$ using data matrix D:

$$\phi : o \longmapsto \phi(o) = (\mathscr{F}(f_1, o), \ldots,\ \mathscr{F}(f_m, o)),\ \mathscr{F}(f_j,\ o) = \begin{cases} 1 & f_j \in o \\ 0 & otherwise \end{cases} \tag{1}$$

For example, in the different tasks of text mining, text documents are represented using words as features (so-called, bag-of-words). Therefore, the rows of matrix D correspond to documents in the collection and the columns correspond to words. Each word is represented using a high dimensional sparse feature vector. Therefore,

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

previous researches (Kriegel *et al.* 2009; Assent 2012) have noted the need for feature engineering using semantic kernels.

The main idea of kernel methods is to map the data objects into a more comparable feature space. The classical representation of objects provided by (1) ignores any semantic relation between features. One of the most important issues is to improve the representation to ensure that objects containing semantically related features are mapped to similar feature vectors. In order to address the challenge of neglecting semantic content of the features in vector space, a transformation of the type $\tilde{\phi}(o) = \phi(o) K_m$ is required, where K_m is a semantic matrix. $\tilde{\phi}(o)$ is a less sparse vector with non-zero entries for all features that are semantically similar to those presented in the data object o. Using this transformation, the corresponding kernel representation takes the form of (2).

$$\tilde{k}\left(o_i,\ o_j\right) = \phi(o_i) K_m \acute{K}_m \phi(\acute{o}_j) = \tilde{\phi}(o_i)\tilde{\phi}(\acute{o}_j) \tag{2}$$

Different choices of the matrix K_m lead to different variants of vector space semantic kernels (Shawe-Taylor and Cristianini 2004). As described in Section 2, building the matrix K_m maybe explicit such as Wikipedia Semantic Kernel (Wang and Domeniconi 2008) or implicit such as Latent Semantic Kernel (Cristianini *et al.* 2002). There are a lot of applications in data mining and machine learning which have a prior knowledge about the relations between features. For the sake of simplicity, the relations between semantic features are modeled as an undirected graph, named *feature similarity graph*.

Definition 2: A *feature similarity graph* for a feature space f is an undirected graph $G = (V, E)$, where vertices are the features in f, while a link between them represents a basic similarity between the corresponding features. Two semantically related features f_i and f_j are connected with an unweighted edge in G. For example, while representing text documents with words as features, WordNet (Finlayson 2014) provides a feature similarity graph which defines different relations between words.

A clique in an undirected graph is a subset of its vertices ($C \subseteq V$) as if every two vertices in the subset are connected by an edge (a complete subgraph) (Mihalcea and Radev 2011). Every clique corresponds to an independent set in the complement graph. A maximal clique is a clique that cannot be extended by including one more adjacent vertex, that is, a clique which does not exist exclusively within the vertex set of a larger clique. A maximal clique in the feature similarity graph represents a subset of highly related features which is not a part of a larger clique. The independence between cliques is reminiscent of the discrimination attribute between features in machine learning (Assent 2012). The main idea of this paper is that these cliques can play an important role as features in comparing complex data objects (*cliques as features*). This means that two objects can be seen similar even if they do not share any features, but the features they contain are elements of common cliques. Accordingly, we can define a semantic matrix (K_m) which maps each concept to the corresponding cliques in the feature similarity graph.

Definition 3: A *clique-based semantic kernel matrix* (K) over a feature similarity graph $G = (V, E)$ is a $m \times l$ matrix, where m is the number of semantic features and l is the number of maximal cliques in G. Each row of K maps a semantic

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

feature to the corresponding maximal cliques in the feature similarity graph. This is a binary matrix where (i, j)th entry gives the existence of feature f_i in the vertex set of clique c_j.

The columns of this matrix correspond to maximal cliques (c_1, c_2, …, c_l) and the rows correspond to features (f_1, f_2,…, f_m). Each entry ($K_{ij} \in \{0, 1\}$) represents the existence of a feature (f_i) in a clique c_j as follows:

$$K = \begin{bmatrix} k_{11} & \cdots & k_{1l} \\ \vdots & \ddots & \vdots \\ k_{m1} & \cdots & k_{ml} \end{bmatrix} \tag{3}$$

A general feature (f_i) is usually found in a lot of cliques and does not give enough weight to other meaningful features. It is possible to apply normalization filter to the original kernel matrix which is inspired by Inverse Document Frequency in classical information retrieval (Turney and Pantel 2010) as follows:

$$K_{i*} = K_{i*} \times \log\left(\frac{l}{1 + |c_j : f_i \in c_j|}\right) \tag{4}$$

Where K_{i*} represents a row in the kernel matrix which corresponds to feature f_i. l is the total number of maximal cliques and $|c_j : f_i \in c_j|$ is the number of cliques where the feature f_i appears. If a feature does not appear in any cliques, this will lead to a division-by-zero. Therefore, the denominator of (4) has been adjusted to $1 + |c_j : f_i \in c_j|$.

We can now define a semantic kernel which leverages the semantic matrix K to map each data object to the corresponding relevant cliques in the feature similarity graph. A *Clique-based Semantic Kernel* represents a data object o by $\tilde{\phi}(o) \longmapsto \phi(o)K$, where K is a clique-based semantic matrix. The new representation model of clique-based semantic kernel $\tilde{\phi}(o)$ is less sparse than the original $\phi(o)$ that has non-zero entries for all semantic features that are semantically related to those present in $\phi(o)$. Equation 5 represents this process using matrix multiplication:

$$F_{nl} = D_{nm} \times K_{ml} \tag{5}$$

Matrix F is the generated feature matrix based on both features and the corresponding cliques in the feature similarity graph. The columns of the feature matrix (F) correspond to maximal cliques (c_1, c_2, …, c_l) and the rows correspond to data objects (o_1, o_2,…, o_n). Each entry (f_{ij}) represents the relationship between clique c_j and object o_i. Finally, the new generated feature vector corresponding to object o_i is:

$$\overrightarrow{F_i} = (f_{i1}, \ldots, f_{il}) \tag{6}$$

A common clique (c_j) appears in a lot of objects and does not give enough weight to other meaningful cliques. So the original feature matrix need a normalization like the kernel matrix in (4). We applied a normalization filter to the original feature matrix which is inspired by Inverse Document Frequency in classical information retrieval (Turney and Pantel 2010):

$$F_{*j} = F_{*j} \times \log\left(\frac{n}{1 + |o_i : c_j \in \mathrm{o_i}|}\right) \tag{7}$$

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

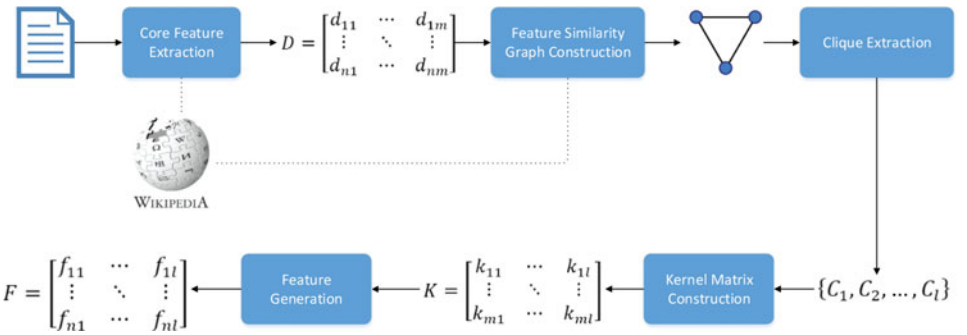


Fig. 1. (Colour online) The overall flow of the proposed clique-based semantic kernel in the task of text relatedness.

Where n is the total number of objects, F_{*j} corresponds to the jth maximal clique (c_j) and $|o_i{:}c_j \in o_i|$ refers to the number of objects where the clique c_j appears. Finally, it's common to use cosine similarity (Turney and Pantel 2010) to find the semantic relatedness of two objects o_i and o_j using the corresponding feature vectors $\overrightarrow{F_i}$ and $\overrightarrow{F_j}$:

$$sim\left(o_i,\ o_j\right) = \cos\left(\overrightarrow{F_i},\ \overrightarrow{F_j}\right) = \frac{\overrightarrow{F_i}.\ \overrightarrow{F_j}}{\left|\overrightarrow{F_i}\right|\left|\overrightarrow{F_j}\right|} = \frac{\sum_{k=1}^{l} f_{ik} * f_{jk}}{\sum_{k=1}^{l} f_{ik}^2\ \sum_{k=1}^{l} f_{jk}^2} \tag{8}$$

In the following section, the effectiveness of the proposed clique-based semantic kernel is evaluated in the application of semantic relatedness between texts and words (Zesch and Gurevych 2010; Zhang *et al.* 2013).

4 Using clique-based semantic kernel for computing text relatedness

The proposed clique-based semantic kernel is crucial in any application of machine learning which has a prior knowledge about the relatedness between features. On the other hand, emerging knowledge repositories in the field of natural language understanding, especially those that are based on collaborative knowledge repositories (Gurevych and Wolf 2010; Gurevych and Zesch 2013; Hovy, Navigli and Ponzetto 2013) such as Wikipedia (Medelyan *et al.* 2009), DBpedia (Bizer *et al.* 2009) and BabelNet (Navigli and Ponzetto 2012), provides a valuable prior knowledge about conceptual features (Jadidinejad and Mahmoudi 2014). In this section, Wikipedia articles and the corresponding hyperlink graph are employed as a conceptual feature repository and it is also tried to validate the usefulness of the proposed clique-based semantic kernel in the field of computing relatedness between text documents.

Figure 1 shows the overall function of our experiments. At the first step, the corpus is mined to extract Wikipedia entities. This process is known as entity linking (Hachey *et al.* 2013) in general, or Wikification (Csomai and Mihalcea 2008) when the target knowledge base is Wikipedia. Entity extraction module provides a data matrix (D) which represents the existence of a specific concept in different documents (bag-of-concepts). After that, it is necessary to leverage the hyperlink structure in Wikipedia to mine the similarity between conceptual features. The

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

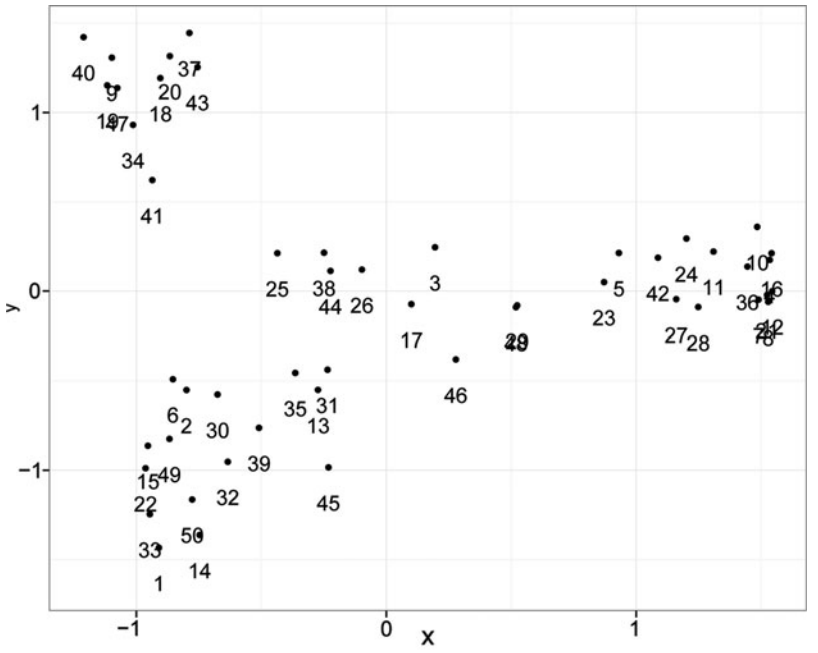


Fig. 2. Visualization of 50 documents of Lee data set in a 2-dimensional space using multidimensional scaling.

output is a structured feature space which is defined as feature similarity graph in Section 3. Clique extraction module derives all the maximal cliques from the feature similarity graph. Each clique corresponds to a coherent subset of related conceptual features. These cliques shape the kernel matrix (K) which relates each clique to its ingredients. Finally, the new feature matrix (F) can be computed by multiplying the data matrix (D) with the kernel matrix (K).

We leverage a collection of 50 text documents from the Australian Broadcasting Corporation's news mail service (Lee, Pincombe and Welsh 2005). This corpus is well analyzed in the application of semantic relatedness (Zhang *et al.* 2013) using different document representation models (Gabrilovich and Markovitch 2009; Tsatsaronis, Varlamis and Vazirgiannis 2010; Taieb, Aouicha and Hamadou 2013; Yazdani and Popescu-Belis 2013). On the other hand, the limited number of text documents allows us to easily analyze different clique-based approaches in a moderate hardware configuration. Following the literature on text semantic relatedness, these documents are paired in all possible ways and compared to the average human judgments using Pearson's product-moment correlation coefficient. 2D visualization of this data set is presented in Figure 2. Each point in the figure corresponds to a document. The position of each document represents the relative distance from others. Obviously, this data set contains four conceptual clusters.

Entity linking (Hachey *et al.* 2013) is the first step in conceptual natural language processing. Building a precise entity linking module is a hot research topic which is out of the scope of this paper. Fortunately, previous researches (Mihalcea and Csomai 2007; Hachey *et al.* 2013; Milne and Witten 2013) in the field of Wikification

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

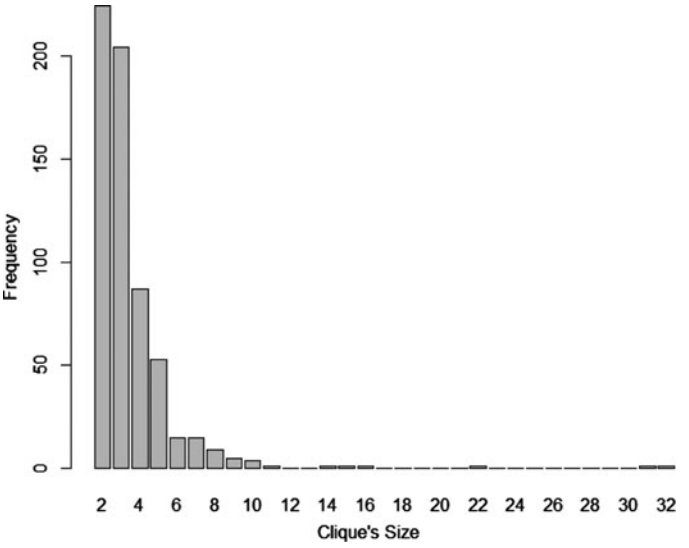


Fig. 3. The number of occurrences of each maximal clique in the feature similarity graph according to cliques size.

are useful in this phase. Wikification is a process in which words of a textual document link to the corresponding Wikipedia Web page (Csomai and Mihalcea 2008). We used Wikifier (Milne and Witten 2013) to map a text fragment to the corresponding concepts in Wikipedia (s_1, s_2, …, s_m). Totally, $m = 496$ unique concept has been detected between 50 documents of Lee data set (Lee *et al.* 2005). For example, consider again two example documents from Lee data set (Lee *et al.* 2005) in Table 1. The most important extracted Wikipedia concepts are shown in bold. According to the provided human judgments, these documents are highly related to each other. Unfortunately, there is no exact match between the extracted concepts.

In order to build the feature similarity graph, a basic similarity measure between articles in Wikipedia is required. Despite different knowledge based and corpus based measures which have been proposed in the literature (Zesch and Gurevych 2010; Zhang *et al.* 2013), the following heuristic rules were leveraged to induce a basic similarity measure to build the feature similarity graph:

(1) If there is a bidirectional hyperlink between two articles u and v, these articles will be adjacent in the corresponding feature similarity graph.
(2) Each article in Wikipedia has at least one parent category (Medelyan *et al.* 2009). If two articles u and v have at least one common parent category (sibling articles), these articles will be adjacent in the corresponding feature similarity graph.

Using the above wiring rules, 2,671 edges have been shaped between $m = 496$ unique concepts. Totally, $l = 623$ maximal cliques have been discovered in the feature similarity graph. The size of the discovered cliques is varying from 2 to 32 vertices. The number of occurrences of each maximal clique is illustrated in Figure 3. This

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

Table 2. *Pearson's correlation of text relatedness scores with human judgements on Lee data set. Correlation is significant at the 0.01 level (2-tailed). Different versions of Clique-based Semantic Kernel (CSK) have been presented*

	r	95% CI
BOW	0.50	0.45–0.54
BOC	0.60	0.56–0.64
CSK-without normalization	0.63	0.59–0.67
CSK-K normalization	0.67	0.63–0.70
CSK-F normalization	0.69	0.66–0.72
CSK-K, F normalization	**0.72**	0.69–0.75
LSA (Lee *et al.* 2005)	0.60	0.56–0.64
ESA (Gabrilovich and Markovitch 2009)	0.72	0.69–0.75

figure reveals that maximal cliques in the graph follow Power-law distribution with long tail (Steyvers and Tenenbaum 2005). This means that the number of cliques with very large size is very low and a large number of cliques have small size. A large clique contains a lot of general concepts related to each other according to the structure of the background knowledge base. For example, a very large clique in our experiments on Wikipedia contains different countries. These cliques are triggered by the existence of each general concept (vertices) in the document. To solve the importance of these large cliques, a weighting formula was presented in (4) and (7).

After extracting different cliques from the features similarity graph, the clique-based semantic matrix (K) can be shaped. The columns of this matrix correspond to maximal cliques ($c_1, c_2, \ldots, c_l$) and the rows correspond to Wikipedia concepts ($s_1, s_2, \ldots, s_m$). Each entry ($K_{ij} \in \{0, 1\}$) represents the existence of a concept in a clique. With the existence of the kernel matrix, it is possible to map documents into the new clique-based feature space using (5). Finally, the relatedness between two documents in the new clique-based feature space are computed using cosine similarity (8). Table 2 shows the Pearson's correlation between the proposed algorithm and human judgment scores in comparison to baselines and the state-of-the-art document representation models (Dumais 2004; Gabrilovich and Markovitch 2009). We leveraged LSA (Dumais 2004) and ESA (Gabrilovich and Markovitch 2009) as well-known representatives of latent and explicit representation models respectively. In the bag-of-words representation model ('BOW'), words appeared in the input document are leveraged as features. On the other hand, in the bag-of-concepts representation model ('BOC'), each document is represented with its concepts (Wikipedia's concepts (Jadidinejad and Mahmoudi 2014) in our experiments). All versions of the proposed feature generation model are significantly better than the baseline bag-of-words and bag-of-concepts. The kernel matrix (K) and the generated feature matrix (F) are normalized according to (4) and (7) respectively. Normalized version of both kernel and feature matrices significantly improves the effectiveness of the original proposed method and yields the best correlation coefficient to the human scores.

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

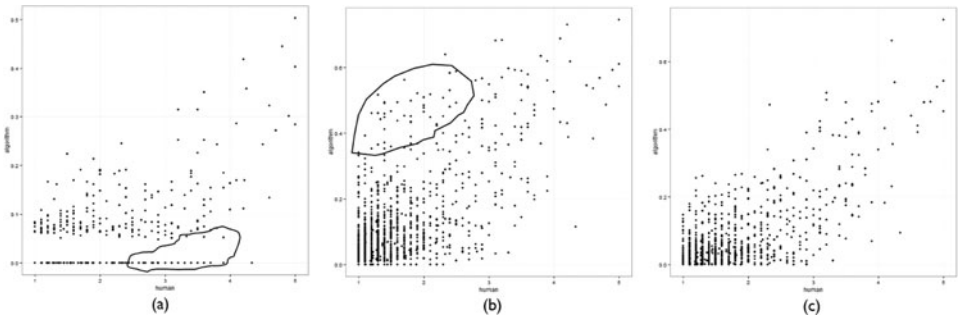


Fig. 4. Comparison of correlation between the algorithm and human scores for different document representation models: (a) bag-of-concepts, (b) clique-based semantic kernel without normalization and (c) clique-based semantic kernel with normalization of both K and F matrices. Each point corresponds to a pair of documents in Lee data set. Horizontal and vertical axis represents the amount of semantic relatedness determined by humans and algorithms respectively. Circled regions in Figures (a) and (b) represent the noisiest decisions in each approach. It is clear that calculated semantic relatedness values in Figure (c) are highly correlated to human decisions.

In order to determine whether the proposed clique-based semantic kernel is statistically significantly superior to baselines, we used G. Zou statistical hypothesis testing (Zou 2007) which represents a testing method for both dependent and independent variables. The advantage of this method is the acknowledgment of the asymmetry of sample distributions for single correlations and it only requires confidence intervals (Singer *et al.* 2013). We used 0.05 and 0.95 for alpha and confidence level respectively. Statistical significance test (Zou 2007) were calculated between the dependent Pearson correlations coefficients produced by the proposed semantic kernel and baselines ('BOW' and 'BOC') . One-tailed hypothesis test was used for assessing the difference between two paired correlations. When comparing the proposed semantic kernel with the baselines and LSA (Lee *et al.* 2005), null hypothesis is rejected. It means that the proposed method is not only statistically significantly better than the baseline, but also better than LSA. On the other hand, null hypothesis is retained when comparing the proposed semantic kernel with ESA (Gabrilovich and Markovitch 2009), therefore the proposed clique-based semantic kernel performs as well as ESA.

Figure 4 shows the correlation between the proposed algorithm and human judgment scores compared with different document representation models presented in Table 2. As shown in Figure 4(a), the baseline bag-of-concept's representation model leads to a lot of vocabulary mismatches between core conceptual features. There are a lot of cases in Figure 4(a), where the human judgment score is high while the baseline algorithm score is zero. For example, two documents presented in Table 1 are highly related according to the provided human judgments ($r = 4.33/5$) but there is no common concept between them. So, the nominator of (8) equals to zero and the bag-of-concept's representation model can't detect the relatedness between them ($r = 0.0$).

On the other hand, the proposed clique-based representation model (Figure 4(b) leverages cliques-of-concepts instead of isolated bag-of-concepts. This method allows

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

semantically related documents to be matched. For example, consider again two documents presented in Table 1, using the proposed clique-based features, these two documents have **24** features (cliques) in common and the overall similarity between them is $r = 0.093$. Some common clique-based features between them are as follows:

- Natural environment: Forest
- Earth:Natural environment: Earth Summit: Environmentalism
- Environmentalism: Sustainability: Earth Summit: Sustainable development
- Environmentalism: Sustainability: Earth Summit: Natural environment

Figure 4(b) reveals that there are some cases which have delusive high scores. Strict analysis of the results have shown that general concepts have high degree of feature similarity graph and naturally appear in a lot of cliques. On the other hand, common cliques appear in a lot of documents and do not give enough weight to other meaningful cliques. This problem is reminiscent of the challenge of general terms in classical information retrieval (Turney and Pantel 2010). Furthermore, (4) and (7) proposed an inverse frequency normalization technique for the kernel matrix (K) and the feature matrix (F) respectively. The effect of the normalization on both kernel and feature matrices is shown in Figure 4(c). As it is clearly illustrated, smoothing general concepts and cliques significantly improves the correlation coefficient ($r = 0.72$ in Table 2).

5 Using clique-based semantic kernel for computing concept similarity

In order to show the applicability of the proposed clique-based semantic kernel in different knowledge bases and applications, in this section WordNet (Finlayson 2014) is leveraged in the task of concept similarity. Unlike Wikipedia, WordNet is an expert-built lexical database. It groups English words into sets of synonyms called'synsets' and records the various semantic relations between these synonym sets. In the following experiments, The MIT Java WordNet Interface (JWI) (Finlayson 2014) has been used to access WordNet programmatically.

Computing semantic similarity in a network of words or concepts plays an important role in different applications of information retrieval and natural language processing (Harispe *et al.* 2013). Assuming similarity as a specific type of relatedness, the goal of concept similarity is to measure the amount of similarity between two concepts using the defined relations in the knowledge base (Schwartz and Gomez 2011).

We leverage a collection of 97 WordNet concept pairs which is a benchmark data set in the task of concept similarity (Schwartz and Gomez 2011), so-called 'ConceptSim' in the following experiments. Schwartz and Gomez (2011) created this data set by mapping each word from Agirre *et al.* word similarity data set (Agirre *et al.* 2009) into WordNet. For example, the corresponding concept pair for the word pair ('tiger', 'jaguar') in Agirre *et al.* word similarity data set (Agirre *et al.* 2009) is ('tiger#n#2', 'jaguar#n#1') in ConceptSim data set (Schwartz and Gomez 2011). The correlation between each algorithmic measure and the gold standard is evaluated using Spearman's rank correlation.

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

ConceptSim contains 152 unique concepts (synsets of WordNet). A subgraph G was built from WordNet which contains 2,796 vertices and 3,087 edges by starting from 152 unique concepts and add all neighbors which are reached by all types of semantic relations. This feature similarity graph contains 2,812 maximal cliques.

Now, we can define a data matrix $D_{n\times m}$ where rows are indexed by the unique concepts of ConceptSim ($n = 152$) and the columns correspond to conceptual features ($m = 2,796$) in the feature similarity graph (G). For example, 'jaguar#n#1' connects to 'big_cat#n#1' and 'Panthera#n#1' in the feature similarity graph (G). So, the row corresponding to 'jaguar#n#1' has two ones among 2,796 entries. Using this matrix, each concept is represented using its neighbors in a high dimensional feature space in which each dimension is associated with a WordNet conceptual feature. The similarity between two concepts c_i and c_j is calculated using cosine measure, so-called '$R_{overlap}$' in the following experiments:

$$sim\left(c_i,\ c_j\right) = \cos\left(\overrightarrow{D_{i*}},\ \overrightarrow{D_{j*}}\right) \tag{9}$$

After that, we can define a clique-based semantic matrix ($K_{m\times l}$) over the feature similarity graph, where $m = 2,796$ is the number of semantic features and $l = 2,812$ is the number of maximal cliques. Each row of K represents a semantic feature with the corresponding maximal cliques in the feature similarity graph (G). This matrix is normalized according to (4).

As described in Section 3, a new clique-based representation can be defined as: $F = D \times K$, where K is a clique-based semantic matrix and D is the original data matrix. This matrix is normalized according to (7). The columns of the feature matrix (F) correspond to maximal cliques and the rows correspond to WordNet concepts.

The new clique-based representation (F) is less sparse than the original representation (D) that has non-zero entries for all related semantic features. The similarity between two concepts c_i and c_j in the new representation is calculated using cosine measure, named 'CSK' in the following experiments:

$$sim\left(c_i,\ c_j\right) = \cos\left(\overrightarrow{F_{i*}},\ \overrightarrow{F_{j*}}\right) \tag{10}$$

Table 3 shows the effectiveness of the proposed clique-based semantic kernel in comparison to the state-of-the-art path based, gloss based and information content based techniques in the field of concept similarity (Budanitsky and Hirst 2006). Experimental results have revealed that the proposed semantic kernel is not only significantly better than baselines, but also comparable with the state-of-the-art results (Patwardhan and Pedersen 2006; Yang and Powers 2006; Resnik 1999).

6 Conclusion

We presented a vector space semantic kernel which leverages concept cliques, a highly coherent subset of concepts, to map high dimensional data objects to a novel feature space where complex data objects can be compared and reduce the vocabulary mismatch and feature sparseness which are main challenges in information retrieval and machine learning respectively. Clique-based semantic

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

Table 3. *Spearman's correlation of concept similarity scores with human judgements on ConceptSim data set. Correlation is significant at the 0.01 level (2-tailed)*

	ρ	95% CI	Type
$R_{overlap}$	0.41	0.23–0.57	Vector-based
$R_{YangPowers}$ (Yang and Powers 2006)	0.63	0.49–0.74	Path-based
$R_{PartwardhanPedersen}$ (Patwardhan and Pedersen 2006)	0.55	0.39–0.67	Gloss-based
S_{Resnik} (Resnik 1999)	0.59	0.45–0.71	IC-based
CSK	0.61	0.45–0.73	Clique-based

kernel is relevant to any applications of machine learning which has a prior knowledge about the relatedness of semantic features. In this paper, there has been a special concentration on the representation of words and texts using WordNet and Wikipedia respectively. Experimental results over a subset of benchmark datasets have revealed the importance of the proposed clique-based semantic kernel in the representation of both words and texts.

Also, the importance of feature weighting has been revealed in our experiments. A simple weighting formula significantly improved the effectiveness of the representation model. One direction in the future work is to follow more sophisticated techniques for feature weighting. Investigating the application of the proposed clique-based semantic kernel in real world problems is another direction of the future work. For example, in bioinformatics, valuable ontologies have been developed and the vocabulary mismatch problem is one of the most important challenges.

Leveraging conceptual ontologies such as DBpedia (Bizer *et al.* 2009) and BabelNet (Navigli and Ponzetto 2012) instead of Wikipedia or WordNet graph is another direction in the future work. These formal ontologies play an important role in the task of semantic interpretation and incorporating these ontologies in the standard representation models of machine learning is a hot research topic. Clique finding is the main drawback of the proposed clique-based semantic kernel when applied to the real world scale applications. Finding all maximal cliques may require exponential time as there are graphs with exponentially many maximal cliques.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. They are also grateful to Dr. Mark Alan Finlayson for supporting of the MIT Java WordNet Interface.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of the Human Language Technologies: The 2009 Annual Conference of the*

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 19–27.

Anderka, M., and Stein, B. 2009. The ESA retrieval model revisited. In *Proceedings of the Thirty-Second international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 670–1.

Assent, I. 2012., Clustering high dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**: 340–350.

Baroni, M., and Lenci, A. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* **36**: 673–721.

Basili, R., Cammisa, M., and Moschitti, A. 2006. A Semantic kernel to classify texts with very few training examples. *Informatica* **30**: 163–172.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. 2009. DBpedia - A crystallization point for the web of data. *Web Semantics* **7**: 154–65.

Bloehdorn, S., Basili, R., Cammisa, M., and Moschitti, A. 2006. Semantic kernels for text classification based on topological measures of feature similarity. In *Proceeding of the Sixth International Conference on Data Mining*, pp. 808–12.

Bloehdorn, S., and Moschitti, A. 2007. Structure and semantics for expressive text kernels. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, ACM, pp. 861–864.

Budanitsky, A., and Hirst, G. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics* **32**: 13–47.

Cristianini, N., Shawe-Taylor, J., and Lodhi, H. 2002. Latent semantic kernels. *Journal of Intelligent Information Systems* **18**: 127–152.

Croce, D., Moschitti, A., and Basili, R. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1034–1046.

Csomai, A., and Mihalcea, R. 2008. Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems* **23**: 34–41.

Dumais, Susan T. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology* **38**: 188–230.

Finlayson, M.A. 2014. Java libraries for accessing the princeton Wordnet: comparison and evaluation. In *Proceedings of the Seventh Global Wordnet Conference*, pp. 713–721.

Fodeh, S., Punch, B., and Tan, P.-N. 2011. On ontology-driven document clustering using core semantic features. *Knowledge and Information Systems* **28**: 395–421.

Gabrilovich, E., and Markovitch, S. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* **34**: 443–498.

Gurevych, I., and Wolf, E. 2010. Expert-built and collaboratively constructed lexical semantic resources. *Language and Linguistics Compass* **4**: 1074–1090.

Gurevych, I., and Zesch, T. 2013. Collective intelligence and language resources: introduction to the special issue on collaboratively constructed language resources. *Language Resources and Evaluation* **47**: 1–7.

Hachey, B., Radford, W., Nothman, J., Honnibal, M. and Curran, J.R. 2013. Evaluating entity linking with Wikipedia. *Artificial Intelligence* **194**: 130–150.

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. 2013. *Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis*. ArXiv e-prints.

Huang, L., Milne, D., Frank, E., and Witten, I. H. 2012. Learning a concept-based document similarity measure. *Journal of the American Society for Information Science and Technology* **63**: 1593–1608.

Hovy, E., Navigli, R., and Ponzetto, S.P. 2013. Collaboratively built semi-structured content and artificial intelligence: the story so far. *Artificial Intelligence* **194**: 2–27.

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

Jadidinejad, A. H., and Mahmoudi, F. 2014. Unsupervised short answer grading using spreading activation over an associative network of concepts. *Canadian Journal of Information and Library Science* **38**: 287–303.

Kriegel, H.-P., Peer, K., and Zimek, A. 2009. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data* **3(1)**: 1:1–1:58.

Lee, M. D., Pincombe, B., and Welsh, M. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, pp. 1254–59.

Liberman, S., and Markovitch, S. 2009. Compact hierarchical explicit semantic representation. In *Proceedings of the IJCAI 2009 Workshop on User-Contributed Knowledge and Artificial Intelligence: An Evolving Synergy (WikiAI09)*, pp. 36–38.

Medelyan, O., Milne, D., Legg, C., and Witten, Ian H. 2009. Mining meaning from Wikipedia. *Int. J. Hum.-Comput. Stud.* **67**: 716–54.

Mehdad, Y., Moschitti, A., and Zanzotto, F. M. 2010. Syntactic/semantic structures for textual entailment recognition. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1020–1028.

Mihalcea, R., and Radev, D. 2011. *Graph-based Natural Language Processing and Information Retrieval*. Cambridge: Cambridge University Press.

Mihalcea, R., and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM conference on Conference on information and knowledge management*, ACM, pp. 233–42.

Milne, D., and Witten, I. H. 2013. An open-source toolkit for mining Wikipedia. *Artificial Intelligence* **194**: 222–239.

Navigli, R., and Ponzetto, S. P. 2012. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* **193**: 217–250.

Patwardhan, S., and Pedersen, T. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of EACL Workshop Making Sense of Sense — Bringing Computational Linguistics and Psycholinguistics TogetherWorkshop Making Sense of Sense—Bringing Computational Linguistics and Psycholinguistics Together*, pp. 1–8.

Resnik, P. 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* **11**: 95–130.

Schwartz, H. A., and Gomez, F. 2011. Evaluating semantic metrics on tasks of concept similarity. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, pp. 299–304.

Shawe-Taylor, J., and Cristianini, N. 2004. *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University.

Singer, P., Niebler, T., Strohmaier, M., and Hotho, A. 2013. Computing semantic relatedness from human navigational paths: a case study on Wikipedia. *International Journal on Semantic Web and Information Systems* **9**: 41–70.

Steyvers, M., and Tenenbaum, J. B. 2005. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science* **29**: 41–78.

Taieb, M. A. H., Aouicha, M. B., and Hamadou, A. B. 2013. Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems* **50**: 260–78.

Tsatsaronis, G., Varlamis, I., and Vazirgiannis, M. 2010. Evaluating entity linking with Wikipedia. *Journal of Artificial Intelligence Research* **37**: 1–40.

Turney, P. D., and Pantel, P. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* **37**: 141–88.

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press

Wang, P., and Domeniconi, C. 2008. Building semantic kernels for text classification using wikipedia. In *Proceedings of the fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 713–721.

Wang, P., Hu, J., Zeng, H.-J. and Chen, Z. 2009. Using Wikipedia knowledge to improve text classification. *Knowledge and Information Systems* **19**: 265–281.

Yang, D., and Powers, D. M. W. 2006. Verb similarity on the taxonomy of Wordnet. *In Proceedings of the third International WordNet Conference (GWC-06)*, pp. 121–128.

Yazdani, M., and Popescu-Belis, A. 2013. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence* **194**: 176–202.

Zesch, T., and Gurevych, I. 2010. Wisdom of crowds versus wisdom of linguists - measuring the semantic relatedness of words. *Natural Language Engineering* **16**: 25–59.

Zhang, Z., Gentile, A. L., and Ciravegna, F. 2013. Recent advances in methods of lexical semantic relatedness a survey. *Natural Language Engineering* **19**: 411–479.

Zou, G. Y. 2007. Toward using confidence intervals to compare correlations. *Psychological Methods* **12**: 399–413.

https://doi.org/10.1017/S135132491500008X Published online by Cambridge University Press