Inf Retrieval (2012) 15:54–92
DOI 10.1007/s10791-011-9172-x

Graph-based term weighting for information retrieval

Roi Blanco · Christina Lioma

Received: 25 August 2010 / Accepted: 30 May 2011 / Published online: 28 June 2011
© Springer Science+Business Media, LLC 2011

Abstract A standard approach to Information Retrieval (IR) is to model text as a bag of words. Alternatively, text can be modelled as a graph, whose vertices represent words, and whose edges represent relations between the words, defined on the basis of any meaningful statistical or linguistic relation. Given such a *text graph*, graph theoretic computations can be applied to measure various properties of the graph, and hence of the text. This work explores the usefulness of such graph-based text representations for IR. Specifically, we propose a principled graph-theoretic approach of (1) computing term weights and (2) integrating discourse aspects into retrieval. Given a text graph, whose vertices denote terms linked by co-occurrence and grammatical modification, we use graph ranking computations (e.g. PageRank Page et al. in The pagerank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998) to derive weights for each vertex, i.e. term weights, which we use to rank documents against queries. We reason that our graph-based term weights do not necessarily need to be normalised by document length (unlike existing term weights) because they are already scaled by their graph-ranking computation. This is a departure from existing IR ranking functions, and we experimentally show that it performs comparably to a tuned ranking baseline, such as BM25 (Robertson et al. in NIST Special Publication 500-236: TREC-4, 1995). In addition, we integrate into ranking graph properties, such as the average path length, or clustering coefficient, which represent different aspects of the topology of the graph, and by extension of the document represented as a graph. Integrating such properties into ranking allows us to consider issues such as discourse coherence, flow and density during retrieval. We experimentally show that this type of ranking performs comparably to BM25, and can even outperform it, across different TREC (Voorhees and Harman in TREC: Experiment and evaluation in information retrieval, MIT Press, 2005) datasets and evaluation measures.

R. Blanco
Computer Science Department, University of A Coruña, A Coruña, Spain
e-mail: rblanco@udc.es

C. Lioma (✉)
Computer Science Department, Stuttgart University, Stuttgart, Germany
e-mail: liomaca@ims.uni-stuttgart.de

Keywords Information retrieval · Graph theory · Natural language processing

1 Introduction

A network is defined as a system of elements that interact or regulate each other. Networks can be mathematically represented as graphs. Typically, the term 'graph' refers to visual representations of the variation of one variable compared to other variables, or the mathematical concept of a set of vertices connected by edges, or data structures based on that mathematical concept; whereas the term 'network' typically refers to interconnected systems of things (inanimate objects or people), or specialised types of the mathematical concept of graphs. Throughout this work, we will use the terms 'graph' and 'network' interchangeably.

Network representations are used widely in many areas of science to model various physical or abstract systems that form web-like interwoven structures, such as the World Wide Web (Web) (Albert et al. 1999), the Internet (Faloutsos et al. 1999), social networks (Barabási et al. 2002; Girvan and Newman 2002;, Krapivsky et al. 2000; Moore and Newman 2000; Newman 2001; Wasserman and Faust 1994), metabolic networks (Feinberg 1980; Jeong et al. 2000; Lemke et al. 2004), food webs (Barrat et al. 2004; McCann et al. 1998; Polis 1998), neural networks (Latora and Marchiori 2003; Sporns 2002; Sporns et al. 2002), transportation networks (Berlow 1999; Guimera et al. 2005; Li and Cai 2004), disease and rumour spreading (Moore and Newman 2000; Pastor-Satorras and Vespignani 2001), or urban infrastructure (Scellato et al. 2005; see Albert 2005; Albert and Barabási 2002; Boccaletti et al. 2006; Christensen and Albert 2007) for overviews). Key to understanding such systems are the mechanisms that determine the topology of their resulting networks. Once the topology of a network is determined, the network can be quantitatively described with measures that capture its most salient properties, by making an analogy between mathematical properties of network topology, such as diameter or density, and real properties of the system being modelled as a network, e.g. Web connectivity or clustering. This allows to draw parallels between the structural mechanics of a physical (or abstract) system and the connectivity of a mathematical object. Based on such parallels, estimations can be made about the system. For instance, in bibliometrics, citation networks are used to estimate the scientific productivity of a field (based on its rate of growth, or its citation preferences Belew 2005) or the importance of an author (based on how many other authors cite him/her, and how important these authors are themselves Newman 2001). Similarly, in Web IR, graph representations of hyperlinked Webpages are used to compute how important a Webpage is, on the basis of how many other Webpages point to it, and how important those Webpages are Page et al. (1998). In such graph theoretic representations of networks, the notion of an 'important' vertex being linked to another vertex is called *recommendation*.

This work models text as a network of associations that connect words (*text graph*). We apply this text graph representation to IR, by deriving graph-based term weights, and by drawing an analogy between topological properties of the graph and discourse properties of the text being modelled as a graph. Text graphs are a well explored area in linguistics (overview in Sect. 2.2). The underlying motivation behind text graphs is that they can represent the non-linear, non-hierarchical structure formation of language in a mathematically tractable formalism. This representation is powerful because it can integrate several aspects of language analysis (topological, statistical, grammatical, or other)

seamlessly into the model. We focus on text graphs that model term co-occurrence and grammatical modification, and we analyse their topology to gain insights into discourse aspects of the text being modelled. We posit that term co-occurrence and grammatical modification reflect language organisation in a subtle manner that can be described in terms of a graph of word interactions. The underlying hypothesis is that in a cohesive text fragment, related words tend to form a network of connections that approximates the model humans build about a given context in the process of discourse understanding (Halliday and Hasan 1976). This position is accepted in linguistics, for instance by Deese's (1965) and Cramer's (1968) earlier work on the relation between word association and the structured patterns of relations that exist among concepts, Hobbs' work on word relationships that 'knit the discourse together', and Firth's (1968b) well-known work about seeking and finding the meaning of words 'in the company they keep'.

Our study of text graphs can be split into two parts:

(a) how the text graph is built
(b) what computations are done on the text graph.

Regarding (a), we build a graph where vertices denote terms, and where edges denote co-occurrence and grammatical relations between terms. Regarding (b), we realise two different types of computations on this graph, aiming either to rank the vertices, or to measure properties of graph topology. The former (vertex ranking) computations allow us to rank each vertex based on properties of the whole graph. The latter (measuring properties of graph topology) allows us to enhance the previously computed term weights with information about topological properties of the graph (and hence of the text), which represent discourse properties of the text.

We apply these term weights and discourse properties to IR in a series of experiments involving building two different types of text graphs, computing four different graph-based term weights, and using these weights to rank documents against queries. We reason that our graph-based term weights do not necessarily need to be normalised by document length (unlike existing term weights) because they are already scaled by their graph-ranking computation. This is a departure from existing IR ranking functions, and we experimentally show that it can perform comparably to established, highly tuned ranking, such as BM25 (Robertson et al. 1995). In addition, we integrate into ranking graph-dependent properties, such as the average path length, or clustering coefficient of the graph. These properties represent different aspects of the topology of the graph, and by extension of the document represented as a graph. Integrating such properties into ranking practically allows us to consider issues such as discourse coherence, flow and density when retrieving documents with respect to queries. These combinations are evaluated on three different Text REtrieval Conference (TREC; Voorhees and Harman 2005) collections (with 350 queries) against a BM25 baseline. We measure retrieval performance using three different standard TREC evaluation measures, and we tune the parameters of all our ranking functions (both the baseline and the graph-based ones) separately for each evaluation measure. In addition, we carry out an extra 'split-train' parameter tuning study, which confirms the stability of our approach across different parameter settings.

There exist numerous uses of graph theory in IR (overviewed in Sect. 2.1). This work contributes an alternative approach, which allows to model term co-occurrence and grammatical relations into retrieval as an integral part of the term weight. In addition, this work contributes a ranking function that contains no document length normalisation and that can perform comparably to a baseline that contains tuned document length normalisation. To our knowledge, this is novel. The final contribution of this work is the analogy

drawn between graph topology properties and aspects of text discourse, which enables us to numerically approximate discourse aspects and successfully integrate them into retrieval. Even though this analogy between graph properties and discourse aspects is not novel in linguistics (see discussion in Sect. 2.2), its implementation into IR is novel, and we experimentally show that it is effective.

The remainder of this paper is organised as follows. Section 2 overviews related work on graph theory in IR (Sect. 2.1), and on graph representations of text (Sect. 2.2). Section 3 introduces some graph theory preliminaries and presents the properties of graph topology used in this work. Section 4 presents the two text graphs we build in this work, and the graph-based term weights that we compute from them. Section 5 presents IR ranking functions that use these graph-based term weights, firstly without normalisation (Sect. 5.1), and secondly enhanced with properties of graph topology (Sect. 5.2). Section 6 describes and discusses the experimental evaluation of these graph-based term weights in IR. Section 7 discusses issues pertaining to the implementation and efficiency of our approach. Section 8 summarises this article and suggests future research directions.

2 Related work

2.1 Graphs in information retrieval

Graph theoretic approaches to IR can be traced back to the early work of Minsky on semantic IR (Minsky 1969), which was followed by several variants of conceptual IR and knowledge-based IR. Numerous variants of graph formalisms have since been used in connectionist[1] approaches to IR (e.g., frame networks, neural networks, spreading activation models, associative networks, conceptual graphs, ontologies; Doszkocs et al. 1990), and numerous approaches have been proposed to shift weights between vertices in the graphs (such as the Inference Network IR model Turtle and Croft 1991 based on the formalism of Bayesian networks Pearl 1988, or Logical Imaging formalisms Crestani and van Rijsbergen 1998). Such connectionist approaches provide a convenient knowledge representation for IR applications in which vertices typically represent IR objects such as keywords, documents, authors, and/or citations, and in which bidirectional links represent their weighted associations or relevance (approximated in terms of semantic or statistical similarity). The propagated learning and search properties of such networks provide the means for identifying relevant information items. One of the main attractions of these models is that, in contrast to more conventional information processing models, connectionist models are 'self-processing' in the sense that no external program operates on the network: the network literally 'processes itself', with 'intelligent behaviour' emerging from the local interactions that occur concurrently between the vertices through their connecting edges.

For instance, one of the earliest neural networks used to model information is the Hopfield net (Hopfield 1982; Hopfield and Tank 1986), in which information was stored in single-layered interconnected neurons (vertices) and weighted synapses (edges). Information was then retrieved based on the network's parallel relaxation method: vertices were activated in parallel and were traversed until the network reached a stable state

[1] The term 'connectionist' has been used to denote most network or graph based approaches, despite the fact that, strictly speaking, classical connectionist systems should consist of weighted, unlabeled links and should exhibit some adaptive learning capabilities.

(convergence). Another early connectionist model explicitly adopted to IR was Belew's AIR (1989), a three-layer neural network of authors, index terms, and documents, which used relevance feedback from users to change the representation of authors, index terms, and documents over time through an iterative learning process. The result was a representation of the consensual meaning of keywords and documents shared by some group of users.

Such connectionist networks have been found to fit well with conventional vector space and probabilistic retrieval models. For instance, Kwok's early three-layer network of queries, index terms, and documents, used a modified Hebbian learning rule to reformulate probabilistic IR (Kwok 1989). Similarly, Wilkinson and Hingston's neural network representations of vector space retrieval used spreading activation through related terms to improve retrieval performance (Wilkinson and Hingston 1991). The above models represent IR applications in terms of their main components of documents, queries, index terms, authors, etc. Network models have also been used in other IR representations, for instance to model the semantic relations between documents as a self-organising Kohonen network (Lin et al. 1991), or to cluster documents (Macleod and Robertson 1991). In addition, similar connectionist approaches have also been used for various classification and optimisation tasks, starting from the early work of Huang and Lippmann (1987).

More recently, the appearance and fast widespread of the Web has caused a resurgence of graph theoretic representations in applications of Web search. Starting with the seminal work of Page et al. (1998) and Kleinberg (1999), the main idea is to draw direct analogies between hypertext connectivity on the Web and vertex connectivity in graphs. Page and Brin proposed the PageRank vertex ranking computation. PageRank uses random walks, which are a way of ranking the salience of a vertex, by taking into account global information recursively computed from the entire graph, rather than relying only on local vertex-specific information. In the context of the Web, where the graph is built out of Webpages (nodes) and their hyper-references (links), PageRank applies a 'random Web surfer' model, where the user jumps from one Webpage to another randomly. The aim is to estimate the probability of the user ending at a given Webpage. There are several alternatives and extensions of PageRank, for instance HITS (Kleinberg 1999), which applies the same idea, but distinguishes the nodes between 'hubs' and 'authorities', where a hub is a Webpage with many outgoing links to authorities, and an authority is a Webpage with many incoming links from hubs. More elaborate ranking algorithms have also been proposed that incorporate information about the node's content into the ranking, for instance anchor text (Chakrabarti et al. 1998), or that involve computationally lighter processes (Lempel and Moran 2001). Such ranking algorithms are used for various tasks, such as Web page clustering (Bekkerman et al. 2007), or document classification (Zhou et al. 2004).

More recently, graph theoretic applications have been used for other applications within IR, for instance IR evaluation measurements (Mizzaro and Robertson 2007), and re-ranking (Zhang et al. 2005). Furthermore, an increasingly popular recent application of graph theoretic approaches to IR is in the context of social or collaborative networks and recommender systems (Craswell and Szummer 2007; Kleinberg 2006; Konstas et al. 2009; Noh et al. 2009; Schenkel et al. 2008).

In the above approaches, the graph is usually built out of the main components of an IR process (e.g. documents and/or queries and/or users). Our work differs because we build the graph out of the individual terms contained in a document. Hence, the object of our representation is not the IR process as such.

2.2 Text as graph

Text can be represented as a graph in various ways, for instance in graphs where vertices denote syllables (Soares et al. 2005), terms (in their raw lexical form, or part-of-speech (POS) tagged, or as senses), or sentences, and where edges denote some meaningful relation between the vertices. This relation can be statistical (e.g. simple co-occurrence Ferrer i and Solé 2001, or collocation[2] Bordag et al. 2003; Dorogovtsev and Mendes 2001; Ferrer i and Solé 2001), syntactic (i Cancho et al. 2007; Ferrer i Cancho et al. 2004; Widdows and Dorow 2002), semantic (Kozareva et al. 2008; Leicht et al. 2006; Motter et al. 2011; Sigman and Cecchi 2002; Steyvers and Tenenbaum 2005), phonological (Vitevitch and Rodrguez 2005), orthographic (Choudhury et al. 2007), discourse (Somasundaran et al. 2009), or cognitive (e.g. free-association relations observed in experiments involving humans Sigman and Cecchi 2002; Steyvers and Tenenbaum 2005). There exist numerous variants of such text graphs. For instance, graphs representing semantic relations between terms can be further subcategorised into *thesaurus* graphs (Leicht et al. 2006; Motter et al. 2011; Sigman and Cecchi 2002; Steyvers and Tenenbaum 2005) and *concept* graphs (Sigman and Cecchi 2002; Steyvers and Tenenbaum 2005). In thesaurus graphs, vertices denote terms, and edges denote sense relations, e.g. synonymy or antonymy. In concept graphs, vertices denote concepts, and edges denote conceptual relations, e.g. hypernymy or hyponymy. Furthermore, in text graphs, the exact definition of the relations that build the graph can vary. For instance, (Mihalcea and Tarau 2004) remove stopwords, and (Hoey 1991) link sentences that share at least two lexically cohesive words. Moreover, edge relations can further combine two or more different statistical, linguistic or other criteria, for instance in syntactic-semantic association graphs (Nastase et al. 2006; Widdows and Dorow 2002). Edge relations can also be further refined, for instance in co-occurrence graphs which define co-occurrence either within a fixed window (Ferrer i and Solé 2001; Masucci and Rodgers 2006; Milo et al. 2004), or within the same sentence (Antiqueira et al. 2009; Caldeira et al. 2005). Optionally, meaningless edge-relations can be filtered out, under any statistical or linguistic interpretation of this (Pado and Lapata 2007). The vertices themselves can be optionally weighted, in line with some statistical or linguistic criterion (e.g. term frequency, or rank in some semantic hierarchy). Such text graphs have been built for several languages, i.e. German, Czech and Romanian (Ferrer i Cancho et al. 2004), Spanish (Vitevitch and Rodrguez 2005), Hindi, Bengali (Choudhury et al. 2007), Japanese (Joyce and Miyake 2008), and even undeciphered Indus script (Sinha et al. 2009).

The applications of such text graphs are numerous. For instance, in the graph representation of the undeciphered Indus script, graph structure is used to detect patterns indicative of syntactic structure manifested as dense clusters of highly frequent vertices (Sinha et al. 2009). Simply speaking, this means being able to detect patterns of syntax in an undeciphered language. Another example can be found in collocation graphs. In such graphs, edges model an idea central to collocation analysis (Sinclair 1991), namely that collocations are manifestations of lexical semantic affinities beyond grammatical restrictions (Halliday and Hasan 1976). Analysing such graphs allows to discover semantically related words based on functions of their co-occurrences (e.g. similarity). A further example is the case of phonological and free-association graphs, which are used to explain

[2] The difference between co-occurrences and collocations is that collocations are significant recurrent co-occurrences (see Sinclair 1991 for more). There exist several measures for distinguishing collocations from insignificant, though recurrent co-occurrences, overviewed in Manning and Schutze (1999).

the human perceptual and cognitive processes in terms of the organisation of the human lexicon (Vitevitch and Rodrguez 2005). The existence of many local clusters in those graphs is seen as a necessary condition of effective associations, while the existence of short paths is linked to fast information search in the brain. Such findings are used to improve navigation methods of intelligent systems (Sigman and Cecchi 2002; Steyvers and Tenenbaum 2005), and also in medicine, where it has been suggested that the effects of disconnecting the most connected vertices of cognitive association graphs can be identified in some language disorders (Motter et al. 2011), and that specific topological properties of these graphs can be quantitatively associated to anomalous structure and organisation of the brain (namely decreased synaptic density) (Ferrer-i-Cancho 2005). Another example of text graph applications can be found in more formal approaches to linguistics, and specifically in transitive reasoning, which uses graph representations of language for logical reasoning. For instance, the Piagetian idea contends that transitive inference is logical reasoning in which relationships between adjacent terms (represented as vertices) figure as premises (Reynal and Brainerd 2005). Yet another example can be found in orthographic association graphs (Choudhury et al. 2007), which study graph topology with the aim to identify aspects of spelling difficulty. In this case, topological properties of the graph are interpreted as difficulties related to spell-checking. Specifically, the average degree of the graph is seen as proportional to the probability of making a spelling mistake, and the average clustering coefficient of the graph is related to the hardness of correcting a spelling error.

More popular applications of text graphs include building and studying dictionary graphs (of semantically-related senses), which are then used for automatic synonym extraction (Blondel et al. 2004; Ho and Fairon 2004; Muller et al. 2006), or word sense disambiguation (Agirre and Soroa 2009; Gaume 2008; Kozima 1993; Mihalcea and Tarau 2004; Véronis and Ide 1990). In such graphs, topological properties are interpreted as indicators of dictionary quality or consistency (Sigman and Cecchi 2002). The popularity of this type of graphs has grown significantly, following the appearance of resources such as WordNet or the Wikipedia, which are themselves network-based and hence explicitly susceptible to graph modelling (see Minkov and Cohen 2008; Pedersen et al. 2004; Widdows and Dorow 2002 for text graph applications using these resources). Another popular and more recent application is opinion graphs, representing opinions or sentiments linked by lexical similarities (Takamura et al. 2007), morphosyntactic similarities (Popescu and Etzioni 2005), term weights like TF-IDF (Goldberg and Zhu 2006), or discourse relations (Somasundaran et al. 2009). Finally, in more mainstream applications of linguistics, text graphs are commonly used to observe the evolution rate and patterns of language (Dorogovtsev and Mendes 2001), while in applications of automatic language processing, text graphs are commonly used to estimate text quality (Antiqueira et al. 2007).

The type of application of text graphs used in this work consists in ranking the graph vertices using random walk computations, like PageRank. Two well-known implementations of random walks on text graphs are TextRank (Mihalcea and Tarau 2004) and LexRank (Erkan and Radev 2004), variants and extensions of which have been applied to keyword detection, word sense disambiguation, text classification (Hassan and Banea 2006), summarisation (by extraction or query-biased Esuli and Sebastiani 2007, or ontology-based Plaza et al. 2008), novelty detection (Gamon 2006), lexical relatedness (Hughes and Ramage 2007), and semantic similarity estimation (Ramage et al. 2009). To our knowledge, the only application of random walks term weights to IR has been our poster study of (Blanco and Lioma 2007), which we extend in this work. Specifically, this work extends (Blanco and Lioma 2007) in three ways (also discussed in Sect. 4). Firstly, in

(Blanco and Lioma 2007), we derived term weights from graphs that modelled solely term co-occurrence. In this work, we compute term weights from graphs that model not only term co-occurrence, but also grammatical modification. We do so using a theoretically principled way in accordance to Jespersen's Rank Theory (Jespersen 1929), a well-known grammatical theory. Secondly, in (Blanco and Lioma 2007), we used graph-based term weights for retrieval by plugging them to the ranking function without considering document length normalisation (we used pivoted document length normalisation at the time). In this work, we look at the issue of document length normalisation very closely, by studying the use of our graph-based term weights in ranking functions without normalisation. Our motivation is that our graph-based term weights do not necessarily need to be normalised by document length because they are already scaled by their graph-ranking computation. Our here-used ranking functions simply combine graph-based term weights with inverse document frequency (*idf*) (Sparck 1972), a well-known statistic of term specificity, and are shown to be comparable to BM25 (Robertson et al. 1995), an established and robust retrieval model that includes an explicit parameterised document length normalisation component. Thirdly, in Blanco and Lioma (2007) we did not consider graph topology at all. In this work, we study topological properties of the text graphs, which we consider analogous to discourse aspects of the text, and we integrate these properties into the ranking process to enhance retrieval.

3 Graph theory preliminaries

3.1 Networks as graphs

Graphs are mathematical representations of networks. The *node* and *link* of a network are referred to as *vertex* (V) and *edge* (E) respectively in graph theory. Formally, an *undirected graph* G is a pair $G = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges, such that $\mathcal{E} \subseteq \mathcal{V}^2$ (see Fig. 1). Edges can be either directed or undirected, as is necessitated by the type of interaction they represent. For instance, the edges of gene-regulatory networks and the Internet are directed, since they depict relationships for which the source and target of the interaction are known; conversely, protein-protein interaction networks and social networks are typically undirected, since these relationships tend to be more mutual (Christensen and Albert 2007). A *directed graph* is a pair $(\mathcal{V}, \mathcal{E})$, where edges point toward and away from vertices. For a vertex $v_i \in \mathcal{V}$, $In(v_i)$ is the set of vertices that point to it and $Out(v_i)$ is the set of vertices that v_i points to, such that:

$$In(v_i) = \{v_j \in \mathcal{V}(v_j, v_i) \in \mathcal{E}\} \tag{1}$$

$$Out(v_i) = \{v_j \in \mathcal{V}(v_i, v_j) \in \mathcal{E}\} \tag{2}$$

$$\mathcal{V}(v_i) = In(v_i) \cup Out(v_i) \tag{3}$$

The *order* of a graph G is the number of its vertices.

3.2 Graph topology properties

Early work on *random graphs* (Erdos and Renyi 1959; Erdos and Renyi 1960; Erdos and Renyi 1961), i.e. on networks for which the connections among nodes have been randomly chosen, pioneered the basic measures of graph topology that would later be extended to the

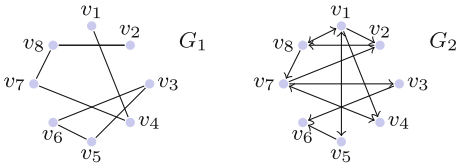

Fig. 1 *Left* an undirected graph (G_1) with eight vertices and seven undirected edges. *Right* a directed graph (G_2) with eight vertices and eleven directed edges

analysis of non-random networks. It was soon established that networks with similar functions have similar graph-theoretical properties (see Albert 2005; Albert and Barabási 2002; Boccaletti et al. 2006; Dorogovtsev and Mendes 2002; Newman 2003 for reviews). Three of the main properties of a graph's topology are its *degree distribution, average path length* and *clustering coefficient*.

3.2.1 Degree distribution

The *degree* $\delta(v_i)$ of a vertex v_i is the number of edges adjacent to v_i. If the directionality of interaction is important (in directed graphs), the degree of a vertex can be broken into an indegree and an outdegree, quantifying the number of incoming and outgoing edges adjacent to the vertex.

The degree of a specific vertex is a local topological measure, and we usually synthesise this local information into a global description of the graph by reporting the degrees of all vertices in the graph in terms of a degree distribution, $P(k)$, which gives the probability that a randomly-selected vertex will have degree k. The degree distribution is obtained by first counting the number of vertices with $k = 1, 2, 3, \ldots$ edges, and then dividing this number by the total number of vertices in the graph (Christensen and Albert 2007). Often the degree distribution is approximated as the *average degree* of a graph, computed as (Mehler 2007):

$$\delta(G) = 2\frac{|\mathcal{E}(G)|}{|\mathcal{V}(G)|} \tag{4}$$

where $\delta(G)$ denotes the average degree of graph G, $|\mathcal{E}(G)|$ denotes the cardinality of edges in G, and $|\mathcal{V}(G)|$ denotes the cardinality of vertices in G.

Extensive work on graph theory in the last decade (reviewed in Albert 2005; Albert and Barabási 2002; Boccaletti et al. 2006; Dorogovtsev and Mendes 2002; Newman 2003) has demonstrated that graphs of similar types of systems tend to have similar degree distributions, and that the vast majority of them have degree distributions that are scale-free (reviewed in Albert and Barabási 2002). The scale-free form of the degree distribution indicates that there is a high diversity of vertex degrees and no typical vertex in the graph that could be used to characterise the rest of the vertices (Christensen and Albert 2007). Practically this means that the degree distribution gives valuable insight into the heterogeneity of vertex interactivity levels within a graph. For instance, in directed graphs, information regarding local organisational schemes can be gleaned by identifying vertices that have only incoming or outgoing edges. These *sinks* and *sources*, respectively, are very likely to have specialised functions. For example, if the graph describes a system of information flow, such as a signal transduction network within a cell, the sources and sinks of the graph will represent the initial and terminal points of the flow. In this case, the

injection points for chemical signals will be sources, and the effectors at which the chemical signals terminate will be sinks (Ma'ayan et al. 2004). Another example is the case of vehicular traffic networks, where sources and sinks take the form of on and offramps in highway systems (Knospe et al. 2002). In these cases, looking at the degree of the respective graph offers insight into how heterogeneous the objects modelled as vertices are, and to what extent some of them (if any) can be considered to be discriminative with respect to the rest.

3.2.2 Average path length

Given an edge adjacent to a vertex, if one traces a path along consecutive distinct edges, only a fraction of the vertices in the graph will be accessible to the starting vertex (Bollobás 1979, 1985). This is often the case in directed graphs, since whether two edges are consecutive depends on their directions. If a path does exist between every pair of vertices in a graph, the graph is said to be *connected* (or *strongly connected* if the graph is directed). The average number of edges in the shortest path between any two vertices in a graph is called *average path length*. Based on this, the *average path length* of the graph $l(G)$ can be estimated as the ratio of the number of its vertices over its degree (Albert and Barabási 2001):

$$l(G) \approx \frac{\ln(|\mathcal{V}(G)|)}{\ln(\delta(G))} \tag{5}$$

where $|\mathcal{V}(G)|$ is the cardinality of vertices in G and $\delta(G)$ is the average degree of G.

For most real networks, the average path length is seen to scale with the natural logarithm of the number of vertices in the graph. In addition, for most real networks, average path length remains small, even if the networks become very large (*small world* property; Watts and Strogatz 1998). Average path length, approximated as the average of inverse distances, can be indicative of the graph's global efficiency, in the sense that it can indicate how long it takes to traverse a graph (Latora and Marchiori 1987, 2003).

3.2.3 Clustering coefficient

The *clustering coefficient* of a vertex measures the proportion of its neighbours that are themselves neighbours. By averaging the clustering coefficients of all vertices in a graph we can obtain an average clustering coefficient of the graph, which is indicative of the strength of connectivity within the graph. Most real networks, including, for example, protein-protein interaction networks, metabolic networks (Wagner and Fell 2001), or collaboration networks in academia and the entertainment industry (Christensen and Albert 2007), exhibit large average clustering coefficients, indicating a high level of redundancy and cohesiveness.

Mathematically, the local clustering coefficient of v_i is given by

$$c(v_i) = \frac{2\mathcal{E}(v_i)}{\delta(v_i)(\delta(v_i) - 1)} \tag{6}$$

where $\mathcal{E}(v_i)$ is the number of edges connecting the immediate neighbours of node v_i, and $\delta(v_i)$ is the degree of node v_i. Alternatively, the average clustering coefficient of a graph can be approximated as the average proportion of the neighbours of vertices that are themselves neighbours in a graph (Albert and Barabási 2001):

$$c(G) \approx \frac{\delta(G)}{|\mathcal{V}(G)|} \tag{7}$$

where $\delta(G)$ denotes the average degree of G, and $|\mathcal{V}(G)|$ denotes the cardinality of vertices in G.

The average clustering coefficient can be used to identify connected graph partitions. By identifying such connected partitions within a network, one may be able to gain a sense of how the network is functionally organised at an intermediate level of complexity. For example, often we come across directed graphs having one or several strongly-connected components. These strongly-connected components are subgraphs whose vertex pairs are connected in both directions. Each strongly-connected component is associated with an in-component (vertices that can reach the strongly-connected component, but that cannot be reached from it) and an out-component (the converse). It has been suggested that the vertices of each of these components share a component-specific task within a given network (Christensen and Albert 2007). For example, in biological networks of cell signal transduction, the vertices of the in-component tend to be involved in ligand-receptor binding, while the vertices of the out-component are responsible for the transcription of target genes and for phenotypic changes (Ma'ayan et al. 2005).

Finally, note that properties of graph topology are connected to each other. For instance, a high average clustering coefficient often indicates a high abundance of triangles (three-vertex cliques) in the graph, which causes short paths to emerge. This has been observed in text graphs across languages, i.e. for German, Czech and Romanian (Ferrer i Cancho et al. 2004).

4 Graph based term weights

This section presents the two different text graphs we build, and the term weights we compute from them. Given a collection of documents, we build text graphs separately for each document. Hence, for the remaining of this article, *text graph* denotes document-based graphs, not collection-based graphs.

4.1 Co-occurrence text graph (undirected)

We represent text as a graph, where vertices correspond to terms, and edges correspond to co-occurrence between the terms. Specifically, edges are drawn between vertices if the vertices co-occur within a 'window' of maximum N terms. The underlying assumption is that all words in the text have some relationship to all other words in the text, modulo window size, outside of which the relationship is not taken into consideration. This approach is statistical, because it links all co-occurring terms, without considering their meaning or function in text. This graph is undirected, because the edges denote that terms simply co-occur, without any further distinction regarding their role. We represent each word in the text as a vertex in the graph. We do not filter out any words, such as stopwords, nor do we focus solely on content-bearing words, such as nouns, when building the graph (unlike Mihalcea and Tarau 2004). An example of such a co-occurrence graph is shown in Fig. 2, which uses a very short text borrowed from (Mihalcea and Tarau 2004), with a window of $N = 4$ terms. The main topological properties of average path length, average degree, and clustering coefficient are also shown.

We derive vertex weights (term weights) from this graph in two different ways:

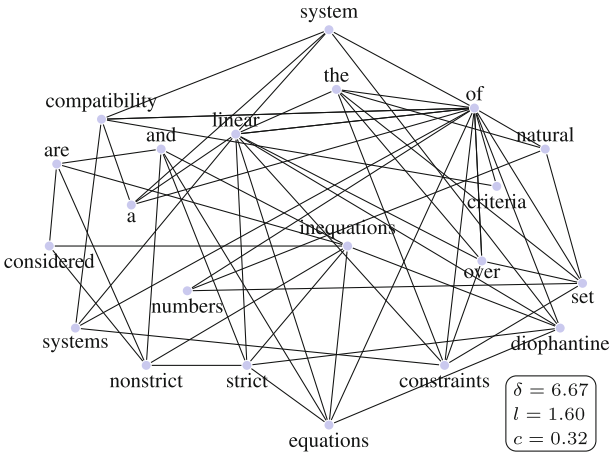

Fig. 2 Co-occurrence graph (undirected) of a short sample text borrowed from (Mihalcea and Tarau 2004). *Vertices* denote terms, and edges denote co-occurrence within a window of 4 terms. Graph topology properties: average degree (δ), average path length (l) and cluster coefficient (c)

- Using a standard graph ranking approach that considers global information from the whole graph when computing term weights (Sect. 4.1.1).
- Using a link-based approach that considers solely local vertex-specific information when computing term weights (Sect. 4.1.2).

4.1.1 Graph ranking weight: textrank

Given a text graph, we set the initial score of each vertex to 1, and run the following ranking algorithm on the graph for several iterations (Mihalcea and Tarau 2004; Page et al. 1998):

$$S(v_i) = (1 - \phi) + \phi \sum_{j \in \mathcal{V}(v_i)} \frac{S(v_j)}{|\mathcal{V}(v_j)|} \quad (0 \leq \phi \leq 1) \tag{8}$$

$S(v_i)$ and $S(v_j)$ denote the score of vertex v_i and v_j respectively, $\mathcal{V}(v_i)$ and $\mathcal{V}(v_j)$ denote the set of vertices connecting with v_i and v_j respectively, and ϕ is a damping factor that integrates into the computation the probability of jumping from a given vertex to another random vertex in the graph. We run (8) iteratively for a maximum number of iterations (100 in this work). Alternatively, the iteration can run until convergence below a threshold is achieved (Mihalcea and Tarau 2004). When (8) converges, a score is associated with each vertex, which represents the importance of the vertex within the graph. This score includes the concept of recommendation: the score of a vertex that is recommended by another highly scoring vertex will be boosted even more. The reasoning behind using such vertex scores as term weights is that: *the higher the number of different words that a given word co-occurs with, and the higher the weight of these words (the more salient they are), the higher the weight of this word.*

Equation (8) implements the 'text surfing model' of Mihalcea and Tarau (2004), which itself is a variation of the original PageRank (Page et al. 1998). The sole difference between (8) and the formula proposed in Mihalcea and Tarau (2004) (and applied to IR in Blanco and Lioma 2007) is that the latter distinguishes between inlinking and outlinking vertices, whereas (8) considers the set of all connected vertices without any sense of

direction. We refer to this weight as *TextRank*, which is the original name used in Mihalcea and Tarau (2004).

4.1.2 Link-based weight: textlink

Given a graph of words, we derive a weight for each vertex directly from the number of its edges:

$$S(v_i) = \delta(v_i) \tag{9}$$

where $\delta(v_i)$ is the average degree of a vertex, as defined in Sect. 3.2.1, Equation 4. The reasoning behind using such vertex scores as term weights is that: *the higher the number of different words that a given word co-occurs with, the higher the weight of this word*. We refer to this weight as *TextLink*.

This weight is a simple approximation, which for this graph is closely linked to term frequency (because in this graph edges are drawn if terms co-occur). This is not necessarily the case for other graphs however, where edges are defined not only on co-occurrence grounds, but also on the basis of grammatical modification, like the graph discussed in the next section.

4.2 Co-occurrence text graph with grammatical constraints (directed)

In our second graph, terms are related according to their co-occurrence as before, but also according to their grammatical modification. Grammatical modification distinguishes a scheme of subordination between words, or otherwise stated, different hierarchical levels. In linguistics these hierarchical levels are called *ranks*, and the special terms *primary, secondary* and *tertiary* refer to the first three ranks, which are typically considered to be semantically more important than the others (Jespersen 1929). Schemes of subordination or different hierarchical levels mean that one word is defined (or modified) by another word, which in its turn may be defined (or modified) by a third word, etc., and that only the word occupying the highest level does not depend on, or does not require the presence of another word. An example follows.

Example 1 `Some (4) furiously (3) speeding (2) cars (1).`

In the above example, the numerals `4, 3, 2, 1` denote the quaternary, tertiary, secondary and primary rank of the words respectively. The primary is typical of nouns, the secondary is typical of adjectives and verbs, the tertiary is typical of adverbs, and the quaternary is typical of the remaining POS, although there can be exceptions. These ranks can be generalised to the case of sentences, for example the primary rank, noun, can be taken to be the subject of a sentence. Under this light, syntactic relations between words imply hierarchical jumps between words. The point to remember is that a word can modify other words of the same or lower rank only. This is one of the principles of Jespersen's Rank Theory (Jespersen 1929), which practically means that a noun can only modify another noun, whereas a verb can modify a noun, verb, or adjective, but not an adverb. It is exactly this modification that we use to build a text graph. This modification has been used in IR applications, for instance in Lioma and Van (2008 and Lioma and Blanco (2009).

Specifically, we build the graph as described in Sect. 4.1, with the difference that we now define vertices pointing to or being pointed to by other vertices (outlinking and

inlinking respectively). A prerequisite for building this graph is to have the text previously grammatically annotated, which we do automatically using a POS tagger.

The resulting graph is directed, where the direction of the edges represents the grammatical modification between terms. For example, given two grammatically dependent terms t_i and t_j, where t_i modifies t_j, we consider the edge direction to be from t_i to t_j, hence t_i points to t_j.

Figure 3 graphically displays an example of such a directed graph for the same short text used in Fig. 2, using the same window of co-occurrence ($N = 4$ terms). Note how the main topological properties of average degree, average path length, and clustering coefficient of this graph differ from the ones computed for the undirected co-occurrence graph in Fig. 2: even though both graphs represent the exact same text, the directed graph has a lower average degree, a higher average path length and a lower clustering coefficient. This is due to the restrictions imposed when drawing edges between vertices, with the practical result of rendering the graph in Fig. 3 less densely linked, longer to traverse, and less clustered.

Having constructed such a graph, we compute the score of a vertex (term weight) in two different ways, similarly to the case of the directed graph, namely by using a graph ranking weight (Sect. 4.2.1), and a link-based weight (Sect. 4.2.2).

4.2.1 Graph ranking weight: PosRank

Since our graph is directed, we compute the original PageRank (Page et al. 1998):

$$S(v_i) = (1 - \phi) + \phi \sum_{j \in In(v_i)} \frac{S(v_j)}{|Out(v_j)|} \quad (0 \leq \phi \leq 1) \tag{10}$$

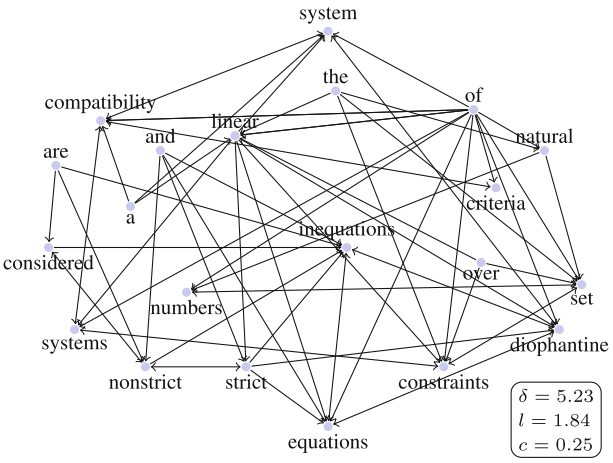

Fig. 3 Co-occurrence graph with grammatical constraints (directed): *vertices* denote POS-tagged terms, the POS of which is ranked from 1 (most salient term) up to 4 (least salient term). Edges denote co-occurrence and grammatical modification between terms. The main topological properties of average degree (δ), average path length (l) and clustering coefficient (c) are lower in value than their equivalent properties computed for Fig. 2, indicating that this graph is less dense, longer to traverse, and less clustered than the undirected graph in Fig. 2

where $In(v_i)$ denotes the set of vertices that modify v_i (in 8 this was simply $\mathcal{V}(v_i)$, i.e. all vertices linking to v_i were considered), and $Out(v_j)$ denotes the set of vertices that v_j modifies (in 8 this was simply $\mathcal{V}(v_j)$, i.e. all vertices linking to v_j were considered). All other notation is as defined for Equation 8. The reasoning behind using such vertex scores as term weights is that: *the higher the number of different words that a given word co-occurs and is grammatically dependent with, and the higher the weight of these words (the more salient they are), the higher the weight of this word*. We refer to this weight as *PosRank*, to stress the fact that it is computed from a graph where words are linked according to their grammatical (or POS) dependence.

4.2.2 Link-based weight: PosLink

Similarly to the case of the undirected graph, we approximate a vertex weight directly from the vertex average degree. However, in order to take into account the aspect of grammatical modification, we consider solely the vertex indegree ($In(v_i)$):

$$S(v_i) = |In(v_i)| \tag{11}$$

In our case, the indegree corresponds to how many words modify a word. The more salient a word (e.g., noun), the higher its indegree. The reasoning behind using such vertex scores as term weights is that: *the higher the number of different words that a given word co-occurs and is grammatically-dependent with, the higher the weight of this word*. We refer to this weight as *PosLink*.

To recapitulate, in Sect. 4 we have presented two ways of representing text as a graph of words, one undirected and one directed. In each case, we have suggested two different ways of assigning scores to vertices (term weights), one based on graph ranking, and one based on links alone. This results in four graph-based term weights, which are summarised in Fig. 4.

5 Graph-based term weights for ranking

We use our four graph-based term weights for retrieval by integrating them to the ranking function that ranks documents with respect to queries. This is a standard use of term weights in IR, and there exist several different ways of doing so, e.g. (Ponte and Croft 1998; Robertson et al. 1995). Most, if not all, of these approaches include an *idf*-like component, which represents the inverse document frequency of a term, defined as the

GRAPH-BASED TERM WEIGHTS OVERVIEW:

TextRank (Equation 8): the higher the number of different words that a given word co-occurs with, and the higher the weight of these words (*recommendation*), the higher the weight of this word.

TextLink (Equation 9): the higher the number of different words that a given word co-occurs with, the higher the weight of this word.

PosRank (Equation 10): the higher the number of different words that a given word co-occurs with and is grammatically related to, and the higher the weight of these words (*recommendation*), the higher the weight of this word.

PosLink (Equation 11): the higher the number of different words that a given word co-occurs with and is grammatically related to, the higher the weight of this word.

Fig. 4 Overview of our four graph-based term weights, with their respective Equation numbers and underlying intuitions

ratio of the total number of documents in a collection over the number of documents that contain a term (Sparck 1972). In addition, most, if not all, of these approaches also include some normalisation component, which re-adjusts the final ranking, typically according to document length, in order to avoid biasing the ranking in favour of longer documents. This re-adjustment, better known as document length normalisation, is an important aspect of ranking, especially for realistic collections that contain documents of varying lengths: without document length normalisation, longer documents tend to score higher since they contain more words and word repetitions (Singhal 2001; Singhal et al. 1996).

More specifically, a typical ranking function estimates the relevance $R(d, q)$ between a document d and a query q as:

$$R(d,q) \approx \sum_{t \in q} w(t,q) \cdot w(t,d) \tag{12}$$

where $w(t, q)$ is the weight of term t in q, and it is usually computed directly from the frequency of the query terms. Since most queries in standard ad-hoc IR are short and keyword-based, the term frequency of each term in the query is 1, hence $w(t,q) = 1$. The second component of (12), $w(t, d)$, is the weight of term t in document d, and, typically, this is the weight that primarily influences the final ranking. There exist various different ways of computing $w(t, d)$, all of which use the frequency of a term in a document (*tf*) one way or another. For example, the BM25 probabilistic ranking function computes $w(t, d)$ as:

$$w(t,d) = w^{(1)} \cdot \frac{(k_3+1) \cdot qtf}{k_3 + qtf} \cdot tfn \tag{13}$$

where $w^{(1)}$ is an *idf*-like component, k_3 is a parameter, *qtf* is the query term frequency, and *tfn* is the normalised term frequency in a document (see Robertson et al. 1995 for their definitions). The normalised term frequency in the document is adjusted according to document length as follows:

$$tfn = \frac{(k_1+1) \cdot tf}{tf + k_1 \cdot \left(1 - b + b \cdot \frac{l}{avl}\right)} \tag{14}$$

where k_1 and b are parameters, and l and avl are the actual and average document length respectively.

Following this practice, we propose two different uses of our graph-based term weights for ranking:

- With *idf*, but without document length normalisation (Sect. 5.1)
- With *idf*, without document length normalisation, but enhanced with properties of graph topology, which represent discourse properties of the text modelled as a graph (Sect. 5.2)

5.1 Raw model (no document length normalisation)

We start off with a general ranking function for estimating the relevance $R(d, q)$ between a document d and a query q, like the one shown in (12). Our goal is to modify the weight of a document with respect to a term ($w(t, d)$) by considering our graph-based term weights. We estimate $w(t, d)$ as:

$$w(t,d) = \log idf \cdot \log tw \tag{15}$$

where *tw* denotes any one of our four proposed graph-based term weights (namely TextRank (Equation 8), TextLink (Equation 9), PosRank (Equation 10), PosLink (Equation 11)).

Equation (15) is very similar to the classical TF-IDF formula (Robertson and Sparck 1976), with the sole difference that we replace term frequency (*tf*) with a graph-based term weight. This replacement is not without grounds: we have found our graph-based term weights to be overall correlated to *tf*, for instance, for the top most relevant document retrieved in experiments with all queries in one of our TREC collections (Disk4&5), Pearson's correlation coefficient is 0.953 between TextRank—*tf*, and 0.714 between PosRank—*tf*.

Equation (15) is reminiscent of the ranking formula we used in our earlier poster work (Blanco and Lioma 2007), however there is an important difference between the two. Whereas in Blanco and Lioma (2007) we applied pivoted document length normalisation to this formula, here we do not apply any sort of normalisation or re-adjustment. We use the 'raw' *idf* and *tw* scores exactly as measured. Doing so with conventional TF-IDF (i.e. applying it without document length normalisation) is detrimental to retrieval. However, in Sect. 6.2 we show that our graph-based term weights can be used without document length normalisation and still perform comparably to BM25 (with tuned document length normalisation) (and outperform normalised TF-IDF).

We refer to Equation15 as our 'raw ranking model', because it contains no normalisation component. The only parameters involved in this ranking function are the window size N of term co-occurrence (for building the text graph) and the damping factor of the iteration for TextRank and PosRank. In Sect. 6.2 we experimentally show the range of N values within which performance is relatively stable. The value of the damping factor ϕ is also typically fixed in literature without major shortcomings (Mihalcea and Tarau 2004; Page et al. 1998).

5.2 Model enhanced with graph topological properties (no document length normalisation)

We present a ranking function that contains three components: (1) an *idf* component, (2) a graph-based term weight component, and (3) a discourse aspect of the document, which is represented as a topological property of the text graph. Specifically, we experiment with three different topological properties, each of which contributes a different discourse aspect into ranking (Sects. 5.2.1–5.2.3). Section 5.2.4 illustrates these graph topological properties for two real sample texts. Section 5.2.5 describes how we integrate these graph topological properties to the ranking formula.

5.2.1 Average degree as a property for ranking documents

The first topological property we use is the average degree of the graph. In Sect. 3.2.1 we discussed how the degree distribution can give valuable insight into the heterogeneity of node interactivity levels within a network. Simply speaking, the higher the average degree of a graph, the more heterogeneous the interactivity levels of its vertices. In our text graph analogy, heterogeneous vertex interaction can be seen as heterogeneous term interaction in a document. For instance, recall Figs. 2 and 3, which represent two different text graphs of the same sample text. The average degree of the first graph is higher than the average degree of the second graph, because the first graph models solely term co-occurrence,

whereas the second graph models term co-occurrence with grammatical modification. Hence, the interaction between the terms is more heterogeneous in the first graph (all co-occurring words are connected) than in the second graph (only co-occurring words that are grammatically dependent are connected).

More generally, in language, words interact in sentences in non-random ways, and allow humans to construct an astronomical variety of sentences from a limited number of discrete units (words). One aspect of this construction process is the co-occurrence and grammatical modification of words, which in our text graph analogy we model as vertex interactions. The more homogeneous these word interactions are, the more cohesive the text is. Cohesiveness is directly linked to discourse understanding, since humans process, understand (and remember) by association (Ruge 1995). More simply, a document that keeps on introducing new concepts without linking them to previous context will be less cohesive and more difficult for humans to understand, than a document that introduces new concepts while also linking them to previous context.

We integrate the average degree of a text graph into ranking with the aim to model the cohesiveness of the document being ranked. Our reasoning is that a more cohesive document is likely to be more focused in its content than a less cohesive document, and hence might make a better candidate for retrieval (this point is illustrated in Sect. 5.2.4). In line with this reasoning, we propose an integration of the average degree of the text graph into ranking, which boosts the retrieval score of lower-degree documents and conversely penalises the retrieval score of higher-degree documents. The exact formula of the integration of the average degree (and of the remaining topological properties we study) is presented at the end of this section (Sect. 5.2.5), so that we do not interrupt the flow of the discussion regarding our reasoning for using properties of graph topology into ranking and their analogy to discourse aspects.

5.2.2 *Average path length as a property for ranking documents*

The second topological property we use is the average path length of the text graph. As discussed in Sect. 2.2, in graphs where edges represent sense relations, shorter path length has been associated to faster information search in the brain. Similarly, in graphs where edges represent term co-occurrence, longer paths result from less term co-occurrence. For example, think of a text graph that has two completely disconnected graph partitions. These partitions correspond to regions in the text that do not share any words at all. If we introduce the same word in both regions of the text, then the graph partitions will become connected and the average path length of the graph will be reduced.

An underlying assumption behind looking at average path length in text graphs that model term co-occurrence and grammatical modification is that the closer two vertices (words) are to each other, the stronger their connection tends to be. This is a generally accepted assumption, supported for instance by studies examining dependence structures derived from the Penn Tree Bank and mapping the probability of dependence to the distance between words, as noted by Gamon (2006) based on Eisner and Smith 2005). Moreover, this combination of distance and co-occurrence measure is also generally accepted (for instance, this can be seen as what *Pointwise Mutual Information* applies on a high-level), and specifically in IR it is reminiscent of the decaying language models proposed by Gao et al. (2005), in the sense that they combine term distance and co-occurrence information into ranking.

We integrate the average path length of a text graph into ranking with the aim to model the discourse dependence of the document being ranked. Our reasoning is that the lower

the average path length of a document, the higher its discourse dependence, or more simply the more tightly knit its discourse is (this point is illustrated in Sect. 5.2.4). In line with this reasoning, we propose an integration of the average path length of the text graph into ranking, which boosts the retrieval score of documents that have lower average path length values, and conversely penalises the retrieval score of documents that have higher average path length values (described in Sect. 5.2.5).

5.2.3 *Clustering coefficient as a property for ranking documents*

In Sect. 2.2 we saw that graph clustering is typically seen as an indicator of the hierarchical organisation of the network being modelled. For instance, in syntactic graphs (i Cancho et al. 2007), clusters are seen as core vocabularies surrounded by more special vocabularies. Similarly, in free-association graphs, clusters are seen as supporter vertices (response words) that gather around one leader vertex (stimulus word), forming a kind of small conceptual community (Jung et al. 2008). In the co-occurrence and grammatical modification graphs used in this work, clustering can be seen as a indication of contextually-bounded 'discourse hubs', in the sense that the clustered terms may very likely share some meaning, which may be more specific, specialised, or contextually-bounded than the general topic(s) of the text. A text graph exhibiting low clustering may indicate a document charasterised by discourse drifting, i.e. a document mentioning several topics in passing, but without having a clear discriminative topic.

We integrate the clustering coefficient of a text graph into ranking with the aim to model the discourse drifting of the document being ranked. Our reasoning is that the higher the clustering coefficient of a document, the more clustered its discourse is with salient and discriminative topics (this point is illustrated in Sect. 5.2.4). In line with this reasoning, we propose an integration of the clustering coefficient of the text graph into ranking, which boosts the retrieval score of documents that have higher clustering coefficient values, and conversely penalises the retrieval score of documents that have lower clustering coefficient values (described in Sect. 5.2.5).

5.2.4 *Illustration*

For the illustration of graph topological properties in text graphs, we consider two different sample texts: a BBC news online article on astronomy and a wikipedia entry on Bill Bryson's book *A Short History of Nearly Everything*. These sample texts, which are freely available online, are displayed in "Appendices 1 and 2" respectively. The BBC article is selected because it discusses a specific topic of a scientific nature in a layman manner. The wikipedia entry is selected because it is discusses a topic-rich book, which covers a multitude of scientific topics also in layman's terms. These two sample texts are of approximately similar size: the BBC text is 537 terms long, and the wikipedia article is 479 terms long.

For each sample text separately, we build two different text graphs: an undirected co-occurrence graph as described in Sect. 4.1, and a directed co-occurrence graph with grammatical constraints as described in Sect. 4.2 (both graphs use a co-occurrence window size of $N = 4$). Figure 5 displays the undirected co-occurrence graph of the BBC sample text, and Fig. 6 displays the undirected co-occurrence graph of the wikipedia sample text.[3] Table 1 displays the topological properties of the two graphs built for each of the two

[3] The illustrations in Figs. 5 and 6 have been generated with CFinder: http://cfinder.org/.

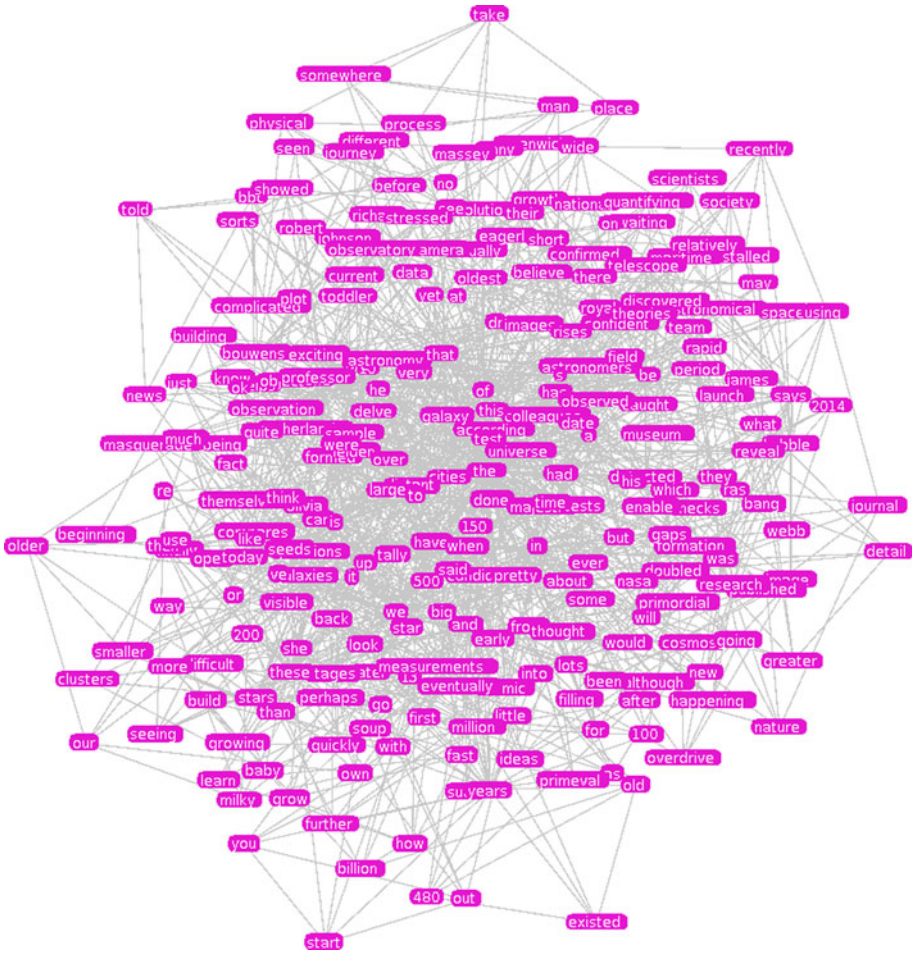

Fig. 5 Undirected co-occurrence text graph for the sample BBC text displayed in "Appendix 1". This graph is built as described in Sect. 4.1, with a co-occurrence window size of $N = 4$. Graph properties are displayed in Table 1

sample texts. Comparing the graph topological properties of the BBC and the wikipedia graphs we observe that the BBC graphs have slightly higher average degree, slightly lower average path length, and slightly higher clustering coefficient, from their respective graphs of the wikipedia text.

Regarding the average degree, we reasoned in Sect. 5.2.1 that it can represent document cohesiveness, and that a document that keeps on introducing new concepts without linking them to previous context will be less cohesive, than a document that introduces new concepts while also linking them to previous context. The higher the average degree of a text graph, the lower the cohesion of the respective text. Indeed, the discourse of the BBC sample text has less cohesion than the discourse of the wikipedia sample text. Specifically, the BBC text discourse shifts across several the third person entities (for instance, `Hubble Space Telescope`, `Professor Richard Bouwens`, `Dr Olivia Johnson`, `he`, `Dr Robert Massey`, `a NASA team`, `the research team`, `astronomers`),

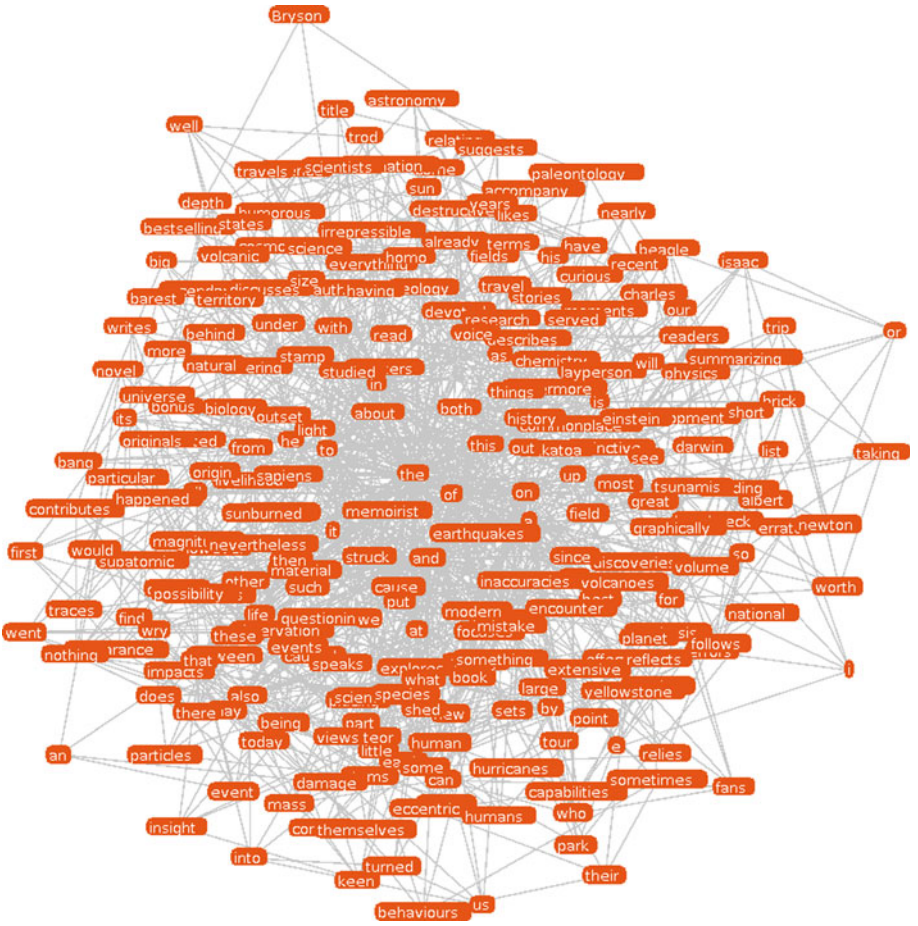

Fig. 6 Undirected co-occurrence text graph for the sample wikipedia entry displayed in "Appendix 2". This graph is built as described in Sect. 4.1, with a co-occurrence window size of $N = 4$. Graph properties are displayed in Table 1

Table 1 Topological properties of the undirected co-occurrence graph (TextGraph) and the directed co-occurrence graph with grammatical constraints (PosGraph) built for the two sample texts (BBC NEWS, WIKIPEDIA)

	TextGraph		PosGraph	
	BBC NEWS	Wikipedia	BBC NEWS	Wikipedia
Average degree	$\delta = 10.32$	$\delta = 9.57$	$\delta = 4.97$	$\delta = 4.69$
Average path length	$l = 1.84$	$l = 1.88$	$l = 2.43$	$l = 2.47$
Clustering coefficient	$c = 0.079$	$c = 0.075$	$c = 0.038$	$c = 0.037$

impersonal structures (for instance, `it is thought, it's very exciting to, there are many`), and repeatedly swaps direct to indirect narration (for instance, `we're seeing, you start out, he compares, we can use`). This is typical of journalistic writing, which generally favours including statements from different agents about

some fact, and referring to those agents in different ways. The wikipedia sample text is more cohesive, because the discourse shifts across fewer entities, mainly Bryson and the book (for instance, `Bryson, Bryson relies, Bryson describes, Bryson also speaks, he states, he the explores, he discusses, he also focuses, this is a book about, the book does`). The discourse is mainly built around Bryson and the book, as opposed to the three different named scientists, and the several other entities of the BBC sample text.

Regarding the average path length, we reasoned in Sect. 5.2.2 that it can represent how tightly knit the discourse is, and that the lower it is, the more tightly knit the discourse is. We observe in Table 1 that the graphs of the BBC sample text have slightly lower average path length than the graphs of the wikipedia sample text. Indeed, the BBC text discusses only the topic of galaxies, and specifically the discovery of the possibly oldest galaxy ever observed. The text does not discuss any other topics, but focuses on different aspects of the discovery and the significance of this discovery. Hence, the discourse of the BBC text is tighly knit on this topic. The wikipedia sample text however discusses a book that covers a wide range of topics pertaining to the history of science across a plethora of fields and aspects (as the title of the book suggests). In this respect, the discourse of the wikipedia sample text is not as tightly knit around a very specific topic.

Finally, regarding the average clustering coefficient, we reasoned in Sect. 5.2.3 that it can represent topical clustering, and that the higher it is, the more clustered the discourse is with salient and discriminative topics (not just passing mentions of topics, i.e. topical drifting, but clusters or *hubs* of topics that are sufficiently discussed). We observe in Table 1 that the graphs of the BBC sample text have a slightly higher clustering coefficient than the graphs of the wikipedia sample text. Indeed, the BBC sample text discusses in more depth the topics it covers (how the oldest galaxy was discovered, what is the significance of this finding), compared to the wikipedia text that simply mentions in passing some of the topics covered in Bryson's book (for instance, `Newton, Einstein, Darwin, Krakatoa, Yellowstone National Park`), without however elaborating enough to create semantic hubs or clusters around these topics.

5.2.5 *Integration into ranking*

We integrate the graph topological properties discussed above into ranking in the same way that query-independent indicators of document retrieval quality are typically integrated into ranking. This use is reminiscent of the approaches presented in Sect. 2, where such graph properties are treated as indicators of thesaurus quality in semantic graphs, for instance. In addition to the above three graph measures, we also use the sum of the graph-based term weights in a graph as a type of graph-equivalent property to document length: in the same way that the sum of the term frequency in a document is seen as an indicator of document length, we sum the graph-based term weights as an indicator of the amount of those weights associated with the nodes of the graph.

We integrate the above graph properties into ranking using the *satu* integration approach, initially suggested in Craswell et al. (2005) for integrating PageRank scores into BM25:

$$w(t,d) = \log idf \cdot \log tw + \psi \frac{P_d}{\kappa + P_d} \tag{16}$$

where tw is any of our four graph-based term weights, P_d is the graph property of document d to be integrated, and ψ, κ are parameters. These parameters can be tuned via extensive 2d

exploration (Craswell et al. 2005), but in our case we fix one and tune the other only (discussed in Sect. 6.1). We use the exact (16) to integrate the estimated clustering coefficient into the ranking ($P_d = c(G)$), but for the remaining three properties we inverse P_d in (16) (i.e. we replace P_d by $1/P_d$) because we assume that the higher these properties are, the lower the relevance ranking, as discussed in Sects. 5.2.1, 5.2.2, 5.2.4. We use the *satu* integration because we wish to integrate into the ranking function graph properties which we assume to be query-independent indicators of the retrieval quality of a document seen as a graph. Other ways of integrating these properties to retrieval are also possible, for instance any of the alternatives presented in Craswell et al. (2005).

6 Experiments

We use our two combinations of ranking with graph-based term weights presented in Sects. 5.1 and 5.2 respectively, in order to match documents to queries. We compare performance against a BM25 baseline.

6.1 Experimental settings

We use the Terrier IR system (Ounis et al. 2007), and extend it to accommodate graph-based computations. For the directed graphs, which require POS tagging, we use the freely available TreeTagger (Schmid 1994) on default settings. We do not filter out stopwords, nor do we stem words, when we build the text graphs and during retrieval. This choice is motivated by findings showing that stemming in general is not consistently beneficial to IR (Harman 1991; Krovetz 2000), and that the overall performance of IR systems is not expected to benefit significantly from stopword removal or stemming (Baeza-Yates and Ribeiro-Neto 1999).

6.1.1 Datasets

For our retrieval experiments we use standard TREC (Voorhees and Harman 2005) settings. Specifically, we use three TREC collections, details of which are displayed in Table 2: Disk4&5 (minus the Congressional Record, as used in TREC), WT2G, and BLOG06. Disk4&5 contains news releases from mostly homogeneous printed media. WT2G consists of crawled pages from the Web. BLOG06 is a crawl of blog feeds and associated documents. These collections belong to different domains (journalistic, everyday Web, blog), differ in size and statistics (Disk4&5 has almost twice as many documents as WT2G, but notably less unique terms than WT2G). For each collection, we use the associated set of TREC queries shown in Table 2. We experiment with short queries (title only), because they are more representative of real Web queries (Ozmutlu et al. 2004). We

Table 2 Dataset features. DISK4&5 exclude the Congressional Record subset, and contain the following: Federal Register (1994), Financial Times (1992–1994), Los Angeles Times (1989–1990)

Collection	Queries	Size (GB)	Documents	Terms	Year
DISK4&5	301–450, 601–700	1.9	528,155	840,536	1989–1994
WT2G	401–450	2	247,491	1,002,586	1997
BLOG06	901–950	25	3,215,171	4,968,020	2005–2006

evaluate retrieval performance in terms of Mean Average Precision (MAP), Precision at 10 (P@10), and binary Preference (BPREF), and report the results of statistical significance testing with respect to the baseline, using the Wilcoxon matched-pairs signed-ranks test.

6.1.2 Parameter tuning

There is one parameter involved in the computation of our graph-based term weights, namely the window size N of term co-occurrence when building the text graph. We vary N within $N = [2, 3, 4, 5, 10, 20, 25, 30]$ and report retrieval performance for each of these values. Note that there is another parameter involved in the computation of TextRank and PosRank only (i.e. the term weights that use recommendation only), namely the damping factor ϕ specified in (8) and (10). We do not tune ϕ, but we set it to $\phi = 0.85$ following (Mihalcea and Tarau 2004; Page et al. 1998).

In addition, there are two parameters involved in the *satu* integration of the graph topological properties into our second ranking function only. The parameters of the *satu* integration are: ψ, κ. We fix $\kappa = 1$ and tune only ψ within $\psi = [1–300]$ in steps of 3. We do not tune both parameters simultaneously, because our aim is to study whether our graph-based term weights are beneficial to retrieval performance, and not to fine-tune their performance in a competitive setting. In spite of this, our graph-based term weights perform very well as shown next, which indicates that the performance reported here may be further improved with full tuning.

Our retrieval baseline, BM25 (Equation 13), includes three tunable parameters: k_1 and k_3, which have little effect on retrieval performance,[4] and b, a document length normalisation parameter. We tune b to optimise retrieval performance by ranging its values within $b = [0–1]$ in steps of 0.05.

6.2 Experimental results

To recapitulate, we have proposed four graph-based term weights: TextRank, TextLink, PosRank, and PosLink. TextRank and TextLink compute term weights from a graph of word co-occurrence; PosRank and PosLink compute term weights from a graph of word co-occurrence and grammatical modification. We use these for ranking without document length normalisation, either 'raw' (i.e. combined only with IDF), or enhanced with the following graph properties: average degree, average path length, clustering coefficient, and the sum of the vertex weights in the graph.

6.2.1 Retrieval precision

Tables 3, 4 and 5 present the retrieval performance of our graph-based term weights and ranking functions against BM25, separately per collection and evaluation measure. The column entitled 'raw' refers to our first 'raw' ranking function; the remaining columns refer to our second ranking function that is enhanced with the graph property mentioned in the header. Specifically, '+Degree, +Path, +Cl. coef, +Sum' refer respectively to the average degree, average path length, clustering coefficient, and sum of graph-based term weights. Note that +Sum is not the sum of the graph properties (i.e. it is not a summation over +Degree and +Path for instance), but the sum of the graph-based term weights of a document, which we use here a graph-based equivalent to document length (this point is

[4] See http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=baselines.

Table 3 Mean average precision (*MAP*) of retrieval results of our ranking with our four graph-based term weights (*TextRank, TextLink, PosRank, PosLink*) compared to baseline ranking with *BM25* (*TFIDF* is displayed for reference). *Raw* denotes ranking without graph topological properties. *+Degree, +Path, +Cl. coef., +Sum* denotes ranking with the respective graph topological properties. Bold font marks MAP $\geq$ baseline. Large font marks best overall MAP, and * marks statistical significance at $p < 0.05$ with respect to the baseline. All scores are tuned as specified in Sect. 6.1.2

Mean average precision (MAP)							
	Raw	+Degree	+Path	+Cl. coef.	+Sum	BM25	TFIDF
WT2G							
TextRank	**0.3033**	**0.3064**	**0.3057**	**0.3083**	**0.3053**	0.2998	0.2268
TextLink	0.2777*	**0.3023**	0.2976	0.2962	0.2850		
PosRank	0.2950	**0.3040**	0.2978	0.2971	0.2942		
PosLink	0.2802	**0.3127**	0.2917	0.2928	0.2894		
D4&5							
TextRank	0.2243	0.2304	0.2243	**0.2329***	**0.2307**	0.2298	0.1935
TextLink	0.2030	0.2233	0.2147*	0.2197*	0.2180*		
PosRank	0.2191	0.2256	0.2205	0.2289	0.2255		
PosLink	0.2020	0.2239	0.2124*	0.2172*	0.2178*		
BLOG							
TextRank	0.3503	0.3501	0.3617	0.3583	0.3531	0.3662	0.2963
TextLink	0.3657	**0.3697**	**0.3947**	**0.3906***	**0.3819**		
PosRank	**0.3874**	**0.3897***	**0.3944***	**0.3918***	**0.3903***		
PosLink	**0.3674**	**0.3903***	**0.3778**	**0.3833***	**0.3833**		

explained in Sect. 5.2.5). The baseline is the BM25 scores (TF-IDF (Robertson and Sparck 1976) scores are also included for reference,[5] but we use BM25, which performs much better than TF-IDF, as a baseline). Tables 3, 4 and 5 display the best scores for each model, after tuning.

Regarding our first 'raw' ranking function, we see that all four of our weights are comparable to the baseline, for all collections and evaluation measures. By comparable, we mean that their performance is between -0.081 and $+0.048$ from the baseline, hence we are not looking at any significant loss or gain in retrieval performance. This is note-worthy, considering the fact the baseline is tuned with respect to document length, whereas our 'raw' ranking is not tuned with that respect. The most gains in retrieval performance associated to our 'raw' ranking are noted with BLOG06, for which our graph-based weights outperform the baseline for MAP (0.3947), P@10 (0.7160), and BPREF (0.4551).

The term weights of the directed graphs (PosRank and PosLink) do not seem to make a significant contribution to retrieval performance in comparison to the term weights of the undirected graphs (TextRank and TextLink). Similar findings have also been reported in other tasks when using graph-based term weights, for instance in keyword extraction (Mihalcea and Tarau 2004), where non-directed graphs fetch higher F-measures than directed graphs. Overall, our graph-based weights in the 'raw' ranking function perform consistently across collections, apart from TextLink and PosLink, which underperform in Disk4&5. A possible reason may be the fact that the Disk4&5 collection contains less unique terms in proportion to document number than the other collections, which implies

[5] TF-IDF is used here with pivoted document length normalisation (Singhal et al. 1996).

Table 4 Precision at retrieved results (*P@10*) of our ranking with our four graph-based term weights (*TextRank, TextLink, PosRank, PosLink*) compared to baseline ranking with *BM25* (*TFIDF* is displayed for reference). *Raw* denotes ranking without graph topological properties. *+Degree, +Path, +Cl. coef., +Sum* denotes ranking with the respective graph topological properties. Bold font marks P@10 ≥ baseline. Large font marks best overall P@10, and * marks statistical significance at $p < 0.05$ with respect to the baseline. All scores are tuned as specified in Sect. 6.1.2

Precision at 10 (P@10)							
	Raw	+Degree	+Path	+Cl. coef.	+Sum	BM25	TFIDF
WT2G							
TextRank	0.4820	**0.4960**	**0.5000**	**0.5000**	**0.4980**	0.4960	0.4120
TextLink	0.4260	0.4840	0.4660	0.4700	0.4400		
PosRank	0.4820	**0.5020**	**0.5020**	**0.5080**	**0.4960**		
PosLink	0.4320	**0.5040**	0.4580	0.4540	0.4500		
D4&5							
TextRank	0.4060	0.4241	0.4112	0.4229*	0.4165	0.4329	0.3855
TextLink	0.3518*	0.4076	0.3896*	0.3940	0.3771*		
PosRank	0.4000	0.4100	0.4020	0.4124	**0.5000***		
PosLink	0.3522	0.4020	0.3831*	0.3847*	0.3755*		
BLOG							
TextRank	0.6380	0.6420*	**0.6680**	0.6560	0.6420	0.6680	0.6000
TextLink	0.6320	0.6500*	**0.6680***	0.6500*	0.6340		
PosRank	**0.7160**	**0.7100**	**0.7140**	**0.7060***	**0.6960***		
PosLink	0.6360	**0.6940**	0.6400	0.6400	0.6440		

high term repetition. However, when we build graphs from text, we link co-occurring terms only once, no matter how many times they actually co-occur. This affects TextLink and PosLink, because these weights rely solely on the average degree of the graph (i.e. the links between terms).

Our second type of ranking that is enhanced with graph topological properties (columns '+Degree, +Path, +Cl. coef., +Sum') performs better than the 'raw' ranking function, and also comparably to BM25. In fact, at all times, the best overall score per collection and evaluation measure is one of our enhanced graph-based term weights. This indicates that the discourse aspects that we integrated into ranking can benefit retrieval performance, not only by bringing in more relevant documents in lower precision ranks, but also by re-ranking the top ranks of the retrieved documents (as shown in the improvements to the P@10 measure). All graph properties seem to work overall equally well, without any significant differences in their measured retrieval performance.

Note that most of our graph-based weights in Tables 3, 4 and 5 are statistically significant with respect to TF-IDF, but only few of them with respect to BM25 (marked * in the Tables).

6.2.2 Parameter sensitivity

The performance of our graph-based term weights depends to an extent on the value of the window size *N* of term co-occurrence. Specifically, the value of the co-occurrence window *N* affects how the text graph is built (which edges are drawn), and hence it is critical to the computation of the graph-based term weights. Figure 7 plots *N* against retrieval

Table 5 Binary Preference (*BPREF*) of the retrieved results of our ranking with our four graph-based term weights (*TextRank, TextLink, PosRank, PosLink*) compared to baseline ranking with *BM25* (*TFIDF* is displayed for reference). *Raw* denotes ranking without graph topological properties. *+Degree, +Path, +Cl. coef., +Sum* denotes ranking with the respective graph topological properties. Bold font marks BPREF ≥ baseline. Large font marks best overall BPREF, and * marks statistical significance at $p < 0.05$ with respect to the baseline. All scores are tuned as specified in Sect. 6.1.2

Binary Preference (BPREF)							
	Raw	+Degree	+Path	+Cl. coef.	+Sum	BM25	TFIDF
WT2G							
TextRank	**0.2963**	**0.3009**	**0.3055**	**0.3043**	**0.2981**	0.2948	0.2338
TextLink	0.2757*	**0.3000**	**0.2994**	**0.2980**	0.2882		
PosRank	0.2926	**0.2961**	**0.2955**	0.2921	0.2920		
PosLink	0.2795	**0.3081**	0.2932	0.2929	0.2884		
D4&5							
TextRank	0.2331	0.2375	0.2329	0.2404*	0.2376	0.2405	0.2094
TextLink	0.2147	0.2303*	0.2246*	0.2277*	0.2275*		
PosRank	0.2319	0.2396	0.2345*	**0.2412**	0.2385		
PosLink	0.2147	0.2315	0.2234*	0.2261*	0.2274*		
BLOG							
TextRank	0.4032	0.4028	0.4104	0.4071	0.4041	0.4161	0.3638
TextLink	**0.4298**	**0.4300**	**0.4469***	**0.4404***	**0.4354**		
PosRank	**0.4477**	**0.4511***	**0.4551**	**0.4493***	**0.4492***		
PosLink	**0.4311**	**0.4489***	**0.4391**	**0.4389**	**0.4404**		

performance, in order to check the stability of the latter across the range of the former. For brevity we show the plots of our first 'raw' ranking only, but we can report that the same behaviour holds for our second enhanced ranking that includes graph properties. In Fig. 7 we observe that performance is overall consistent across collection, evaluation measure, and graph-based term weight. N values between 5 and 30 seem to perform better, and we can confirm that this also holds for our enhanced ranking in these collections. Among the N values between 5 and 30 that perform well, $N = 10$ performs well in terms of MAP and BPREF. A window value of $N = 10$ can be practically interpreted as the subsentence context that our approach considers when weighting term salience. Given that the average sentence length for English is 20 words (Sigurd et al. 2004), this choice of context $N = 10$ seems reasonable; in fact, $N = 10$ has been used in other text processing tasks that consider context within sentences, e.g. the context-based word sense disambiguation of Schütze and Pedersen (Schütze and Pedersen 1995).

Regarding our second enhanced ranking, the performance reported in Tables 3, 4 and 5 depends to an extent on the value of parameter ψ, used when integrating the topological graph properties into ranking. Specifically, the value of ψ controls the influence of the graph property upon the final weight. To this end, we conduct additional experiments in order to test the parameter stability of our graph-based term weights, in a split-train scenario. We focus on our second enhanced ranking function, which contains the integration parameter ψ. We split the Disk4&5 queries into two sets (as split by TREC) and we use queries 301–450 for tuning the parameter ψ, and 601–700 for testing. The aim is to check if the scores reported so far are the product of fine tuning, or if we can expect a relatively similar performance with parameter values that are not necessarily optimal.

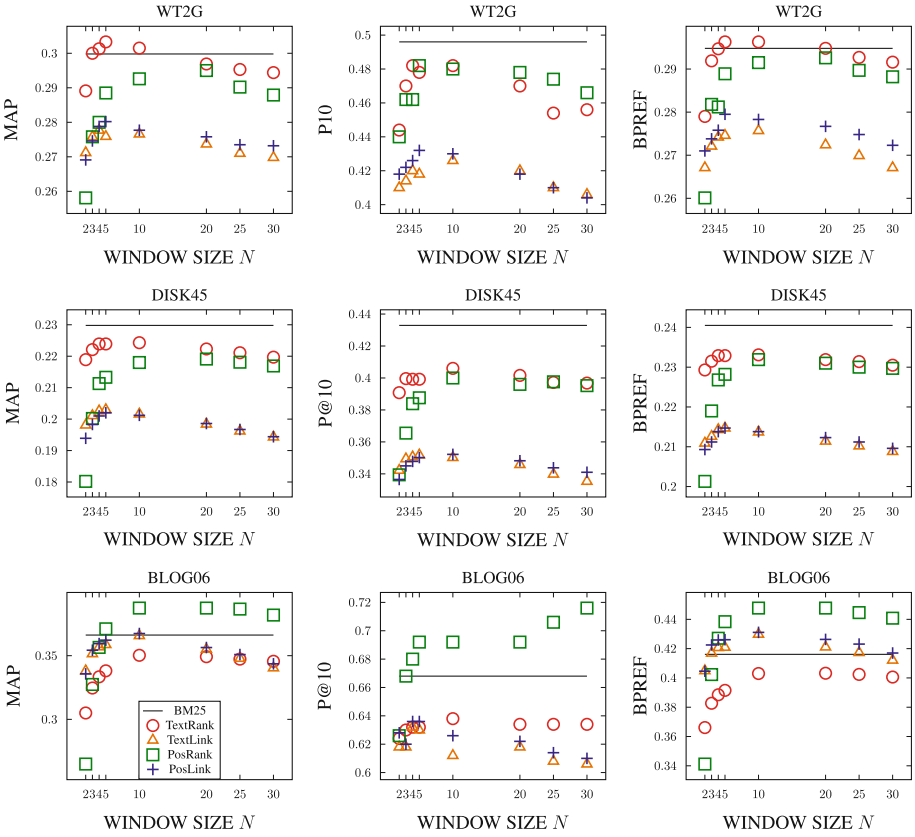

Fig. 7 Retrieval performance (measured in MAP, P@10, BPREF) across the full value range of parameter N (the 'window size' of term co-occurrence)

Table 6 shows retrieval performance for Disk4&5, where we see that the scores obtained by parameters trained on a different query set (column 'train') are not far off the actual best scores (column 'best'). The value of ψ used in these runs is displayed separately in Table 7. We see that at most times the trained value is very close to the best value. These results indicate that the value of ψ can be estimated with relatively little tuning, because it is overall stable across different query sets and evaluation measures.

7 Implementation and efficiency

This section discusses issues pertaining to the implementation and efficiency of our graph-based term weights in an IR system. Graph-based term weights can be computed at indexing time, not at querying time, hence they imply no delay to the IR system's response time to the user. Typically, when a document is fed into the system for indexing, it has to be read from disk for cleaning (extracting the content from Web pages or removing page templates in news articles, for instance), tokenising, parsing etc. It is at that time that the additional overhead of the graph-based term weight computation is applied. Specifically, this overhead is introduced by the algorithm that iterates over (8)–(11) for each of our four

Table 6 Retrieval performance measured in MAP, P@10, BPREF using the TextRank graph-based term weight only on Disk4&5. TextRank is enhanced with four different graph properties (in separate rows). The parameter of this integration is tuned for one subset of the queries, and the tuned parameter values are used with a different subset of the queries ('train' column). Column 'best' reports the actual best performance. There is no notable or significant difference between 'tuned'-'best', indicating that parameter tuning can be ported to different query sets, hence it is potentially robust. The actual parameter values are shown in Table 7

	MAP		P@10		BPREF	
	Train	Best	Train	Best	Train	Best
TextRank—DISK4&5						
Degree	0.2730	0.2743	0.4283	0.4293	0.2578	0.2586
Path	0.2630	0.2630	0.4051	0.4127	0.2486	0.2505
Cl. coef.	0.2779	0.2796	0.4414	0.4414	0.2628	0.2690
Sum	0.2722	0.2741	0.4222	0.4323	0.2585	0.2614
PosRank—DISK4&5						
Degree	0.2680	0.2682	0.4061	0.4182	0.2643	0.2648
Path	0.2573	0.2590	0.3919	0.4111	0.2543	0.2551
Cl. coef.	0.2750	0.2754	0.4202	0.4222	0.2697	0.2712
Sum	0.2688	0.2688	0.4172	0.4192	0.2650	0.2657

Table 7 Parameter ψ values of the integration of graph properties into ranking, as explained in the caption of Table 6

	ψ for best MAP		ψ for best P@10		ψ for best BPREF	
	Train	Best	Train	Best	Train	Best
TextRank—DISK4&5						
Degree	109	190	172	124	109	148
Path	13	13	34	10	13	1
Cl. coef.	16	22	22	22	13	22
Sum	235	298	151	274	208	298
PosRank—DISK4&5						
Degree	118	91	178	100	100	118
Path	28	10	31	4	25	16
Cl. coef.	37	52	31	22	31	40
Sum	298	298	226	286	286	298

graph-based term weights respectively. Figure 8 displays the time overhead in miliseconds associated with the computation of Equation 8, taken as a function of the number of iterations (varied from 0 to 1,000), for computing the TextRank terms weights of the Disk4&5 collection. Figure 8 plots the running time for the different number of iterations averaged over the whole document set, and their variance, computed using an Intel Core 2 Due 3GHz Linux box with 4GB of RAM. We observe that running time increases approximately linearly with the number of iterations, although the overhead is negligible when the number of iterations is less than 20 (<1ms).

The random-walk based approaches as introduced in Sect. 4 approximate an infinite Markov chain on a finite number of steps. Hence, a second question is what would be the

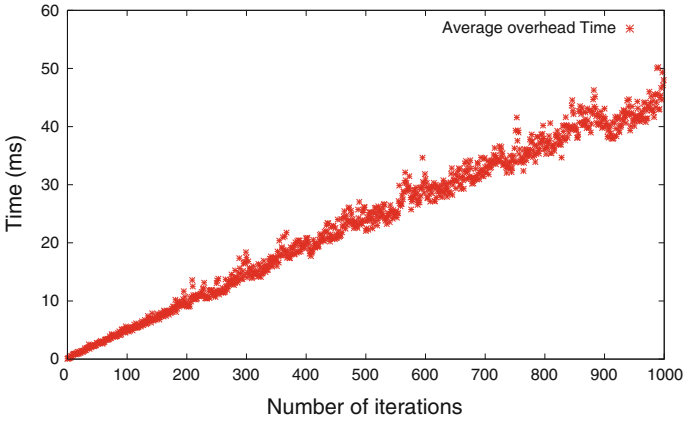

Fig. 8 Overhead in miliseconds introduced by the algorithm. The x-axis represents the number iterations over (8)

required number of iterations of the algorithm to converge. To answer this question, we compare the weights obtained by iterating over (8) a certain number of times vs. iterating a large number of times (ideally, infinite). This provides insights on the number of iterations we need for convergence, or at which point the weights produced by the algorithm are indistinguishable. Specifically, we compute the weights that result after iterating the algorithm in (8) 10K times as ground truth. We then report the mean squared error of the weights that the algorithm produces after iterating for a lower number of times. Figure 9 plots the mean average squared error (MSE) and variance for a given number of iterations. We observe that the MSE decreases exponentially with the number of iterations and that it is close to zero after 100 iterations. In general the error is very low ($<10^{-6}$) after a few iterations (≈ 20). Hence, we can conclude that the algorithm requires a low number of iterations to produce weights that are indistinguishable from those produced using a much higher number of iterations. This operation can be computed offline and independently for every document, thus term weights can be computed efficiently using parallel programming paradigms such as Hadoop.[6]

Overall, the low number of iterations and limited running time implies that our graph-based term weighting algorithm can process documents with a minimum overhead. Additionally, the fact that the whole process can be performed in an offline fashion and completely distributed, allows us to conclude that the processing of these kind of weights could scale up to Web-sized collections.

Using graph-based weights for document ranking has no overhead at query time, compared to other efficient approaches like Ahn and Moffat's impact sorted posting lists (Anh and Moffat 2005). These posting lists store term-information using *impacts*, which are values that represent a score. In Anh and Moffat (2005) an impact is a combination of term frequency and a document length normalisation component. Graph-based term weights would be therefore stored and indexed in the same way, allowing for efficient high-throughput query processing, as fast as other impact-based posting indexing approaches.

[6] http://hadoop.apache.org.

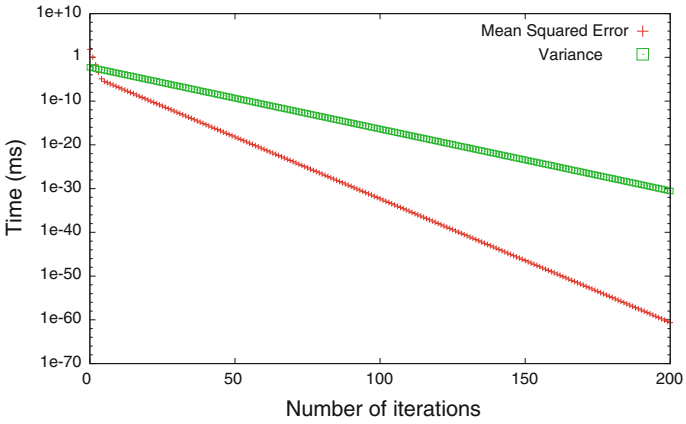

Fig. 9 Average difference between the weights computed at a given number of iterations and the weights computed using 10K iterations, over (8)

8 Conclusions

Starting from a graph representation of text, where nodes denote words, and links denote co-occurrence and grammatical relations, we used graph ranking computations that take into account topological properties of the whole graph to derive term weights, in an extension of the initial proposal of Mihalcea and Tarau (2004). We built two different text graphs, an undirected one (for term co-occurrence) and a directed one (for term co-occurrence and grammatical dependence), and we extracted four graph-based term weights from them. These weights encoded co-occurrence and grammatical information as an integral part of their computation. We used these weights to rank documents against queries without normalising them by document length. In addition, we integrated into ranking graph topological properties (average degree, path length, clustering coefficient). An analogy was made between properties of the graph topology and discourse aspects of the text modelled as a graph (such as cohesiveness or topical drift). Hence, integrating these graph topological properties into ranking practically meant considering discourse aspects of the documents being ranked for retrieval.

Experiments with three TREC datasets showed that our graph-based term weights performed comparably to an established, robust retrieval baseline (BM25), and consistently across different datasets and evaluation measures. Further investigation into the parameters involved in our approach showed that performance was relatively stable for different collections and measures, meaning that parameter training on one collection could be ported to other datasets without significant or notable loss in retrieval performance.

A possible future extension of this work is to use the graph-based term weights presented in this article to rank sentences. This application is reminiscent of the LexRank (Erkan and Radev 2004) approach, which computes sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. Specifically, in LexRank, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences. Sentence similarity is measured based on sentence salience, which in its turn is typically defined in terms of the presence of particular important words or in terms of similarity to a centroid pseudo-sentence. A possible extension of our work could be to use the term salience computed from our

graph-based term weights in order to deduce measures of sentence salience. Such an extension would in fact combine our graph-based term weights to the LexRank approach, and could potentially offer valuable insights into the application of graph-based term weights for sentence extraction.

Further future work includes using other graph ranking computations, e.g. HITS (Kleinberg 1999) to derive term weights, which implies defining vertices as hubs and authorities from a linguistic viewpoint. We also intend to experiment with weighted text graphs, in order to better represent meaningful linguistic relations between words modelled as vertices. This would involve determining how to *learn* the edge weights using training data to perform a weighed or personalized PageRank (Chakrabarti 2007). Our use of such graph-based term weights for IR may be further improved by applying further ranking functions (the two functions presented here had a straightforward TF-IDF format), or re-ranking functions, such as the non-symmetric similarity functions used by Kurland and Lee (2010). Lastly, the encouraging results of the graph topological properties reported here require further investigation, with respect to their integration into ranking and potential combination of more than one property.

Acknowledgments We thank the blind reviewers for their insightful feedback which contributed in strengthening this work.

Appendix 1: Sample text BBC news online article

Sample text from BBC news online[7] (publicly available):

> The Hubble Space Telescope has detected what scientists believe may be the oldest galaxy ever observed. It is thought the galaxy is more than 13 billion years old and existed 480 million years after the Big Bang. A Nasa team says this was a period when galaxy formation in the early Universe was going into overdrive. The image, which has been published in Nature journal, was detected using Hubble's recently installed wide field camera. According to Professor Richard Bouwens of Leiden Observatory in the Netherlands: We're seeing these galaxies—star cities—that are building themselves up over cosmic time. The research team observed rapid growth over a relatively short period of time: Their sample data showed just one galaxy visible about 500 million years after the Big Bang. But this rises to 10 galaxies some 150 million years later. The tally has doubled about 100 million years later. You start out with these little seeds in the very early Universe which would eventually have formed stars, then star clusters, baby galaxies then eventually these large majestic galaxies that we know today, according to Professor Bouwens. It's very exciting to see this complicated physical process actually take place somewhere that no man has seen before, Professor Bouwens told BBC News. He compares the early galaxy to a toddler: It is much smaller than older galaxies like our own Milky Way and it is growing more quickly. We can use these measurements to learn how fast galaxies grow and build up with cosmic time, according to Professor Bouwens. Dr Olivia Johnson of the Royal Greenwich Observatory at the National Maritime Museum says that quantifying the rapid evolution of the Universe will reveal a greater detail about what was happening in the early cosmos—such as when the first stars and galaxies formed. These are big, open questions in astronomy and the fact that we are finally able to look into the primordial universe for the first time is quite exciting, she said.

[7] http://www.bbc.co.uk/news/science-environment-12289840, accessed on 31 January 2011.

The fact that we are finally being able to look into the primordial universe for the first time is quite exciting. Dr Robert Massey of the Royal Astronomical Society (RAS) says the new image from Hubble will enable astronomers to test their current theories of the evolution of the Universe. Professor Bouwens stressed that the observation had yet to be confirmed but that he and his colleagues were pretty confident that they had discovered the oldest galaxy caught on camera to date. There are many different sorts of objects that can masquerade or look very much like these distant objects. We've done lots of checks and lots of tests and we think that this candidate is OK, he said. It's filling in the gaps.Although we have ideas about the formation of the Universe, it is quite difficult to go from the primeval soup in the early stages of the Universe to the Universe we are in. Images like the one we have today helps plot that journey. Astronomers are eagerly awaiting the launch of Nasa's James Webb telescope in 2014 which will be able to delve perhaps 200 million years further back in cosmic time when galaxies were just beginning.

(Term count: 537)

Appendix 2: Sample text wikipedia entry

Sample text from the publicly available wikipedia entry on Bill Bryson's *A Short History of Nearly Everything*[8]:

As the title suggests, bestselling author Bryson (In a Sunburned Country) sets out to put his irrepressible stamp on all things under the sun. As he states at the outset, this is a book about life, the universe and everything, from the Big Bang to the ascendancy of Homo sapiens. *This is a book about how it happened*, the author writes. *In particular how we went from there being nothing at all to there being something, and then how a little of that something turned into us, and also what happened in between and since*. What follows is a brick of a volume summarizing moments both great and curious in the history of science, covering already well-trod territory in the fields of cosmology, astronomy, paleontology, geology, chemistry, physics and so on. Bryson relies on some of the best material in the history of science to have come out in recent years. This is great for Bryson fans, who can encounter this material in its barest essence with the bonus of having it served up in Bryson's distinctive voice. But readers in the field will already have studied this information more in-depth in the originals and may find themselves questioning the point of a breakneck tour of the sciences that contributes nothing novel. Nevertheless, to read Bryson is to travel with a memoirist gifted with wry observation and keen insight that shed new light on things we mistake for commonplace. To accompany the author as he travels with the likes of Charles Darwin on the Beagle, Albert Einstein or Isaac Newton is a trip worth taking for most readers. Bryson describes graphically and in layperson's terms the size of the universe, and that of atoms and subatomic particles. He then explores the history of geology and biology, and traces life from its first appearance to today's modern humans, placing emphasis on the development of the modern Homo sapiens. Furthermore, he discusses the possibility of the Earth being struck by a meteor, and reflects on human capabilities of spotting a meteor before it impacts the Earth, and the extensive damage that such an event would cause. He also focuses on some of the

[8] http://en.wikipedia.org/wiki/A_Short_History_of_Nearly_Everything.

most recent destructive disasters of volcanic origin in the history of our planet, including Krakatoa and Yellowstone National Park. A large part of the book is devoted to relating humorous stories about the scientists behind the research and discoveries and their sometimes eccentric behaviours. Bryson also speaks about modern scientific views on human effects on the Earth's climate and livelihood of other species, and the magnitude of natural disasters such as earthquakes, volcanoes, tsunamis, hurricanes, and the mass extinctions caused by some of these events. The book does however contain some inaccuracies and errors, see Errata (i.e. a list of errors) for *A Short History of Nearly Everything*.

(Term count: 479)

References

Agirre, E., & Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *EACL* (pp. 33–41). The Association for Computer Linguistics.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science, 118*, 4947–4957.

Albert, R., & Barabási, A. L. (2001). Statistical mechanics of complex networks. *CoRR cond-mat/0106096*.

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Review of Modern Physics, 74*, 47–97.

Albert, R., Jeong, H., & Barabási, A. L. (1999). The diameter of the world wide web. *CoRR cond-mat/9907038*.

Allan, J., Aslam, J. A., Sanderson, M., Zhai, C., & Zobel, J. (Eds.). (2009). *Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2009*. Boston, MA, USA: ACM. July 19–23.

Anh, V. N., & Moffat, A. (2005). Simplified similarity scoring using term ranks. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '05* (pp. 226–233). New York, NY, USA: ACM. doi:http://doi.acm.org/10.1145/1076034.1076075.http://doi.acm.org/10.1145/1076034.107607.

Antiqueira, L., Oliveira Jr., O. N., Costa, L. F., & Nunes, M. G. V. (2009). A complex network approach to text summarization. *Information Science, 179*(5), 584–599. doi:http://dx.doi.org/10.1016/j.ins.2008.10.03.

Antiqueira, L. L., Pardo, T. A. S., Nunes, M., & Oliveira, J. O. N. (2007). Some issues on complex networks for author characterization. *Inteligencia Artificial, Revista Iberoamericana de IA, 11*(36), 51–58. url: http://iajournal.aepia.org/aepia/Uploads/36/420.pdf

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). *Modern information retrieval*. New York: ACM Press/Addison-Wesley.

Baeza-Yates, R. A., Ziviani, N., Marchionini, G., Moffat, A., & Tait, J. (Eds.) (2005). *SIGIR 2005: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*. Salvador, Brazil: ACM. August 15–19.

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications, 311*(3-4), 590–614. doi:10.1016/S0378-4371(02)00736-7. http://www.sciencedirect.com/science/article/B6TVG-45S9HG2-1/2/dff30ba73ddd8820aca3e7f072aa788.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of National Academic Science, 101*(11), 3747–3752.

Bekkerman, R., Zilberstein, S., & Allan, J. (2007). Web page clustering using heuristic search in the web graph. In *IJCAI* (pp. 2280–2285).

Belew, R. K. (2011). Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. In Belkin and van Rijsbergen (1989), pp. 11–20.

Belew, R. K. (2005). Scientific impact quantity and quality: Analysis of two sources of bibliographic data. *CoRR abs/cs/0504036*.

Belkin, N. J., & van Rijsbergen, C. J. (Eds.). (1989). *SIGIR'89, 12th international conference on research and development in information retrieval*. Cambridge, Massachusetts, USA: ACM. June 25–28 (Proceedings).

Berlow, E. L. (1999). Strong effects of weak interactions in ecological communities. *Nature, 398*, 330–334.

Blanco, R., & Lioma, C. (2007). Random walk term weighting for information retrieval. In *SIGIR* (pp. 829–830).

Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P., & Dooren, P. V. (2004). A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM RevIEW, 46*(4), 647–666. doi:http://dx.doi.org/10.1137/S003614450241596.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. U. (2006). Complex networks: structure and dynamics. *Physics Reports, 424*, 175–308.

Bollobás, B. (1979). *Graph theory: An introductory course*. New York: Springer.

Bollobás, B. (1985). *Random graphs*. London: Academic Press.

Bookstein, A., Chiaramella, Y., Salton, G., & Raghavan, V. V. (Eds.). (1991). *Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval*. Chicago, Illinois, USA: ACM. October 13–16 (Special Issue of the SIGIR Forum).

Bordag, S., Heyer, G., & Quasthoff, U. (2003). Small worlds of concepts and other principles of semantic search. In T. Bhme, G. Heyer, & H. Unger (Eds.), *IICS, lecture notes in computer science* (Vol. 2877, pp. 10–19). Springer. url: http://dblp.uni-trier.de/db/conf/iics/iics2003.html#BordagHQ0.

Caldeira, S. M. G., Lobao, T. C. P., Andrade, R. F. S., Neme, A., & Miranda, J. G. V. (2005). *The network of concepts in written texts*.

i Cancho, R. F., Capocci, A., & Caldarelli, G. (2007). Spectral methods cluster words of the same class in a syntactic dependency network. *International Journal of Bifurcation and Chaos, 17*(7), 2453–2463.

i Cancho, R. F., Capocci, A., & Caldarelli, G. (2007). Spectral methods cluster words of the same class in a syntactic dependency network. *International Journal of Bifurcation and Chaos, 17*(7), 2453–2463.

Cao, G., Nie, J. Y., & Bai, J. (2005). *Integrating word relationships into language models*. In: R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait (Eds.), *SIGIR* (pp. 298–305).

Chakrabarti, S. (2007). Dynamic personalized pagerank in entity-relation graphs. In *Proceedings of the 16th international conference on World Wide Web, WWW '07* (pp. 571–580). New York, NY, USA: ACM. doi:http://doi.acm.org/10.1145/1242572.124265. URL: http://doi.acm.org/10.1145/1242572.1242650

Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., & Kleinberg, J. M. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks, 30*(1–7), 65–74.

Choudhury, M., Thomas, M., Mukherjee, A., Basu, A., & Ganguly, N. (2007). How difficult is it to develop a perfect spell-checker? A cross-linguistic analysis through complex network approach. In *Proceedings of the second workshop on TextGraphs: Graph-based algorithms for natural language processing* (pp. 81–88). Rochester, NY, USA: Association for Computational Linguistics. url: http://www.aclweb.org/anthology/W/W07/W07-021.

Christensen, C., & Albert, R. (2007). Using graph concepts to understand the organization of complex systems. *International Journal of Bifurcation and Chaos, 17*(7), 2201–2214.

Cramer, P. (1968). *Word association*. New York, USA: Academic Press.

Craswell, N., Robertson, S. E., Zaragoza, H., & Taylor, M. J. (2005). Relevance weighting for query independent evidence. In *SIGIR* (pp. 416–423).

Craswell, N., & Szummer, M. (2007) Random walks on the click graph. In Kraaij et al. (2007), pp. 239–246.

Crestani, F., & van Rijsbergen, C. J. (1998). A study of probability kinematics in information retrieval. *ACM Transaction of Informational System, 16*(3), 225–255.

Deese, J. (1965). *The structure of associations in language and thought*. Baltimore, USA: The John Hopkins Press.

Dorogovtsev, S. N., & Mendes, J. F. F. (2001). Language as an evolving word web. *Proceedings of The Royal Society of London. Series B, Biological Sciences 268*(1485), 2603–2606. doi:10.1098/rspb.2001.1824. url: http://www.isrl.uiuc.edu/amag/langev/paper/dorogovtsev01languageAs.htm.

Dorogovtsev, S. N., & Mendes, J. F. F. (2002). Evolution of networks. *Advances in Physics, 51*, 1079–1187. doi:10.1098/rspb.2001.1824.
http://www.isrl.uiuc.edu/amag/langev/paper/dorogovtsev01languageAs.htm.

Doszkocs, T. E., Reggia, J., & Lin, X. (1990). Connectionist models and information retrieval. *Annual Review of Information Science and Technology (ARIST), 25*, 209–260.

Eisner, J., Smith, N. A. (2005). Parsing with soft and hard constraints on dependency length. In *Proceedings of the international workshop on parsing technologies (IWPT)* (pp. 30–41). Vancouver. http://cs.jhu.edu/jason/papers/#iwpt05.

Erdos, P., & Renyi, A. (1959). On random graphs i. *Publicationes Mathematicae (Debrecen), 6*, 290–297.

Erdos, P., & Renyi, A. (1960). On the evolution of random graphs. *Publication Mathematical Institution of Hungarian Academic Science, 5*, 17–61.

Erdos, P., & Renyi, A. (1961). On the evolution of random graphs. *Bulletin Institution of International Statistics, 38*, 343–347.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR), 22*, 457–479.

Esuli, A., & Sebastiani, F. (2007) Pageranking wordnet synsets: An application to opinion mining. In *The Association for Computer Linguistics (ACL)*.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the internet topology. In *SIGCOMM* (pp. 251–262).

Feinberg, M. (1980). Chemical oscillations, multiple equilibria, and reaction network structure. In W. Stewart, W. Rey, & C. Conley (Eds.), *Dynamics of reactive systems* (pp. 59–130). New York: Academic Press.

Ferrer i Cancho, R. (2005). The structure of syntactic dependency networks: Insights from recent advances in network theory. In G. Altmann, V. Levickij, & V. Perebyinis (Eds.), *The problems of quantitative linguistics* (pp. 60–75). Chernivtsi: Ruta.

Ferrer i Cancho, R., & Solé, R. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics, 8*(3), 165–173.

Ferrer i Cancho, R., Solé, R. V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physics Review E, 69*(5), 051–915. doi:10.1103/PhysRevE.69.051915.

Firth, J. R. (1968b). A synopsis of linguistic theory. In F. R. Palmer (Ed.), *Selected papers of J.R. Firth 1952–1959* (pp. 168–205). London: Longmans.

Gamon, M. (2006) Graph-based text representation for novelty detection. In *Proceedings of TextGraphs: The first workshop on graph based methods for natural language processing* (pp. 17–24). New York City: Association for Computational Linguistics. url:http://www.aclweb.org/anthology/W/W06/W06-380.

Gaume, B. (2008). Mapping the forms of meaning in small worlds. *International Journal of Intelligence System, 23*(7), 848–862.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of National Academic Science USA, 99*(12), 7821–7826.

Goldberg, A., Zhu, X. (2006). Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of TextGraphs: The first workshop on graph based methods for natural language processing* (pp. 45–52). New York City: Association for Computational Linguistics. url:http://www.aclweb.org/anthology/W/W06/W06-380.

Guimera, R., Mossa, S., Turtschi, A., & Amaral, L. A. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of National Academic Science USA, 102*, 7794–7799.

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Harman, D. (1991). How effective is suffixing? *JASIS, 42*(1), 7–15.

Hassan, S., Banea, C. (2006). Random-walk term weighting for improved text classification. In *Proceedings of TextGraphs: The first workshop on graph based methods for natural language processing* (pp. 53–60). New York City: Association for Computational Linguistics. url:http://www.aclweb.org/anthology/W/W06/W06-380.

Ho, N. D., & Fairon, C. (2004). Lexical similarity based on quantity of information exchanged—synonym extraction. In *RIVF* (pp. 193–198).

Hoey, M. (1991). *Patterns of lexis in text*. Oxford, UK: Oxford University Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*(8), 2554–2558.

Hopfield, J. J., & Tank, D. W. (1986). Computing with neural circuits: A model. *Science, 233*, 625–633.

Huang, W. Y., & Lippmann, R. (1987). *Neural net and traditional classifiers*. In D. Z. Anderson (Ed.) *NIPS* (pp. 387–396). American Institue of Physics.

Hughes, T., & Ramage, D. (2007). Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 581–589). Prague, Czech Republic: Association for Computational Linguistics. url: http://www.aclweb.org/anthology/D/D07/D07-106.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabsi, A. L. (2000). The large-scale organization of metabolic networks. *Nature, 407*(6804), 651–654. doi: 10.1038/35036627 url:http://dx.doi.org/10.1038/3503662.

Jespersen, O. (1929). *The philosophy of grammar*. London: Allen and Unwin.

Joyce, T., & Miyake, M. (2008). Capturing the structures in association knowledge: Application of network analyses to large-scale databases of japanese word associations. In T. Tokunaga, A. Ortega (Eds.), *Lecture notes in computer science (LKR)* (Vol. 4938, pp. 116–131). Springer.

Jung, J., Makoshi, N., & Akama, H. (2008). Associative language learning support applying graph clustering for vocabulary learning and improving associative ability. In *ICALT* (pp. 228–232). IEEE.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of ACM, 46*(5), 604–632.
Kleinberg, J. M. (2006). Social networks, incentives, and search. In: E. N. Efthimiadis, S. T. Dumais, D. Hawking, & K. Järvelin (Eds.), *SIGIR* (pp. 210–211). ACM.
Knospe, W., Santen, L., Schadschneider, A., Schreckenberg, M. (2002). Single vehicle data of highway traffic: Microscopic description of traffic phases. *Physical Review, E65*, 056133.
Konstas, I., Stathopoulos, V., & Jose, J. M. (2009). *On social networks and collaborative recommendation*. In Allan et al. (2009), pp. 195–202.
Kozareva, Z., Riloff, E., & Hovy, E. (2008). Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of ACL-08: HLT* (pp. 1048–1056). Columbus, Ohio: Association for Computational Linguistics. url: http://www.aclweb.org/anthology/P/P08/P08-111.
Kozima, H. (1993). Similarity between words computed by spreading activation on an english dictionary. In *EACL* (pp. 232–239).
Kraaij, W., de Vries, A. P., Clarke, C. L. A., Fuhr, N., & Kando, N. (Eds.). (2007). *SIGIR 2007: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*. Amsterdam, The Netherlands: ACM. July 23–27.
Krapivsky, P. L., Redner, S., & Leyvraz, F. (2000). Connectivity of growing random networks. *Physical Review Letters, 85*, 4629–4632.
Krovetz, R. (2000). Viewing morphology as an inference process. *Artificial Intelligence, 118*(1–2), 277–294.
Kurland, O., & Lee, L. (2010). Pagerank without hyperlinks: Structural re-ranking using links induced by language models. In Baeza-Yates et al. (2005), pp. 306–313.
Kwok, K. L. (2011) A neural network for probabilistic information retrieval. In Belkin and van Rijsbergen (1989), pp. 21–30.
Latora, V., & Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical Review Letters, 87*, 198701–198704.
Latora, V., & Marchiori, M. (2003). Economic small-world behaviour in weighted networks. *European Physics Journal, B32*, 249–263.
Leicht, E. A., Holme, P., & Newman, M. E. J. (2006) Vertex similarity in networks. *Physical Review E*, (73).
Lemke, N., Herédia, F., Barcellos, C. K., dos Reis, A. N., & Mombach, J. C. M. (2004). Essentiality and damage in metabolic networks. *Bioinformatics, 20*(1), 115–119.
Lempel, R., & Moran, S. (2001). SALSA: The stochastic approach for link-structure analysis. *ACM Transaction on Informational System, 19*(2), 131–160.
Li, W., & Cai, X. (2004). Statistical analysis of airport network of china. *Physical Review, E69*, 046106.
Lin, X., Soergel, D., Marchionini, G. *A self-organizing semantic map for information retrieval*. In Bookstein et al. (1991), pp. 262–269.
Lioma, C., & Blanco, R. (2009). Part of speech based term weighting for information retrieval. In: M. Boughanem, C. Berrut, J. Mothe, & C. Soulé-Dupuy (Eds.), *ECIR, lecture notes in computer science* (Vol. 5478, pp. 412–423). Springer.
Lioma, C., & Van Rijsbergen, C. J. K. (2008). Part of speech n-grams and information retrieval. *RFLA, 8*, 9–22.
Ma'ayan, A., Blitzer, R. D., & Iyengar, R. (2004). Toward predictive models of mammalian cells. *Annual Review of Giophysics and Biomolecular Structure*, 319–349.
Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, et al. (2005). Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science, 309*(5737), 1078–1083.
Macleod, K. J., & Robertson, W. (1991). A neural algorithm for document clustering. *Information Processing & Management, 27*(4), 337–346.
Manning, C. D., & Schutze, H. (1999). *Foundations of statistical language processing*. London: The MIT Press.
Masucci, A. P., & Rodgers, G. J. (2006). Network properties of written human language. *Physics Review E, 74*(2), 026–102. doi:10.1103/PhysRevE.74.026102.
McCann, K., Hastings, A., & Huxel, G. R. (1998). Weak trophic interactions and the balance of nature. *Nature, 395*, 794–798.
Mehler, A. (2007). Large text networks as an object of corpus linguistic studies. In: *Corpus linguistics. An international handbook of the science of language and society*.
Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. In *EMNLP* (pp. 404–411).
Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., et al. (2004). Superfamilies of evolved and designed networks. *Science, 303*(5663), 1538–1542. url:http://dx.doi.org/10.1126/science.108916.

Minkov, E., & Cohen, W. W. (2008). *Learning graph walk based similarity measures for parsed text*. In *EMNLP* (pp. 907–916). ACL.

Minsky, M. L. (1969). *Semantic information processing*. Cambridge: The MIT Press.

Mizzaro, S., & Robertson, S. (2007). *Hits hits trec: exploring ir evaluation results with network analysis*. In Kraaij et al. (2007), pp. 479–486

Moore, C., & Newman, M. E. J. (2000). Epidemics and percolation in small-world networks. *Physical Review, E61*, 5678–5682.

Motter, A. E., de Moura, A. P. S., Lai, Y. C., & Dasgupta, P. (2011). Topology of the conceptual network of language. *Physics Review E, 65*(6).

Muller, P., Hathout, N., & Gaume, B. (2006). Synonym extraction using a semantic distance on a dictionary. In *Proceedings of TextGraphs: The first workshop on graph based methods for natural language processing* (pp. 65–72). New York City: Association for Computational Linguistics. url: http://www.aclweb.org/anthology/W/W06/W06-3811

Nastase, V., Sayyad-Shirabad, J., Sokolova, M., & Szpakowicz, S. (2006). Learning noun-modifier semantic relations with corpus-based and wordnet-based features. In *AAAI*. AAAI Press

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the USA, 98*(2), 404–409. doi:10.1073/pnas.021544898 url:http://dx.doi.org/10.1073/pnas.02154489.

Newman, M. E. J. (2003). The structure and function of complex networks. *Siam Review, 45*, 167–256.

Noh, T. G., Park, S. B., Yoon, H. G., Lee, S. J., & Park, S. Y. (2009). An automatic translation of tags for multimedia contents using folksonomy networks. In Allan et al. (2009), pp. 492–499.

Ounis, I., Lioma, C., Macdonald, C., & Plachouras, V. (2007). *Research directions in terrier: A search engine for advanced retrieval on the Web*. Novatica/UPGRADE Special Issue on Web Information Access.

Ozmutlu, S., Spink, A., & Ozmutlu, H. C. (2004). A day in the life of Web searching: An exploratory study. *Information Processing & Management, 40*(2), 319–345.

Pado, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics, 33*(2), 161–199.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project. url: citeseer.ist.psu.edu/page98pagerank.html.

Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physics Review Letter, 86*(14), 3200–3203. doi:10.1103/PhysRevLett.86.3200.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). *Wordnet: Similarity—Measuring the relatedness of concepts*. In D. L. McGuinness, & G. Ferguson (Eds.) *AAAI* (pp. 1024–1025). AAAI Press/The MIT Press.

Plaza, L., Daz, A., Gervs, P. (2008). Concept-graph based biomedical automatic summarization using ontologies. In *Coling 2008: Proceedings of the 3rd textgraphs workshop on graph-based algorithms for natural language processing* (pp. 53–56). Manchester, UK: Coling 2008 Organizing Committee. url:http://www.aclweb.org/anthology/W08-200.

Polis, G. A. (1998). Ecology: Stability is woven by complex webs. *Nature, 395*, 744–745.

Ponte, J. M., & Croft, W. B. (1998). *A language modeling approach to information retrieval*. In *SIGIR* (pp. 275–281). ACM.

Popescu, A. M., & Etzioni, O. (2005) Extracting product features and opinions from reviews. In *HLT/EMNLP*. The Association for Computational Linguistics.

Ramage, D., Rafferty, A. N., & Manning, C. D. (2009). Random walks for text semantic similarity. In *Proceedings of the 2009 workshop on graph-based methods for natural language processing (TextGraphs-4)* (pp. 23–31). Suntec, Singapore: Association for Computational Linguistics. url:http://www.aclweb.org/anthology/W/W09/W09-3204

Reynal, V. F., & Brainerd, C. J. (2005). Fuzzy processing in transitivity development. *Annals of Operations Research, 23*(1), 37–63.

Robertson, S., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society of Information Science, 27*, 129–146.

Robertson, S., Walker, S., Beaulieu, M., Gatford, M., & Payne, A. (1995). Okapi at trec-4. In *NIST Special Publication 500-236: TREC-4*.

Ruge, G. (1995). Human memory models and term association. In Fox, E. A., Ingwersen, P., Fidel, R. (Eds.), *SIGIR* (pp. 219–227). ACM Press.

Scellato, S., Cardillo, A., Latora, V., & Porta, S. (2005). The backbone of a city. *European Physics Journal B, 50*(physics/0511063. 1–2), 221–225 (manuscript not submitted to the proceedings NEXT-SigmaPhi).

Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J. X., et al. (2008). Efficient top-k querying over social-tagging networks. In S. H. Myaeng, D. W. Oard, F. Sebastiani, T. S. Chua, & M. K. Leong (Eds.), *SIGIR* (pp. 523–530). ACM.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International conference on new methods in language processing* (pp. 44–49).

Schütze, H., & Pedersen, J. O. (1995). Information retrieval based on word senses. In *Symposium on document analysis and information retrieval* (pp. 161–175).

Sigman, M., & Cecchi, G. A. (2002). Global organization of the WordNet lexicon. *Proceedings of the National Academy of Sciences 3*(99), 1742–1747.

Sigurd, B., Eeg-Olofsson, M., van de Weijer, J., Eeg-Olofsson, M., & van de Weijer, J. (2004). Word length, sentence length and frequency: Zipf's law revisited. *Studia Linguistica, 58*(1), 37–52.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineer Bulletin, 24*(4), 35–43.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In: H. P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), *SIGIR* (pp. 21–29). ACM.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *SIGIR* (pp. 21–29)

Sinha, S., Pan, R. K., Yadav, N., Vahia, M., & Mahadevan, I. (2009). Network analysis reveals structure indicative of syntax in the corpus of undeciphered indus civilization inscriptions. In: *Proceedings of the 2009 workshop on graph-based methods for natural language processing (TextGraphs-4)* (pp. 5–13). Suntec, Singapore: Association for Computational Linguistics. url:http://www.aclweb.org/anthology/W/W09/W09-3202

Soares, M. M., Corso, C., & Lucena, L. S. (2005). Network of syllables in portuguese. *Physica A: Statistical Mechanics and its Applications, 355*(2–4), 678–684. doi:10.1016/j.physa.2005.03.017. url:http://www.isrl.uiuc.edu/amag/langev/paper/soares05networkOfSyllables.htm.

Somasundaran, S., Namata, G., Getoor, L., & Wiebe, J. (2009). Opinion graphs for polarity and discourse classification. In: *Proceedings of the 2009 workshop on graph-based methods for natural language processing (TextGraphs-4)* (pp. 66–74). Suntec, Singapore: Association for Computational Linguistics. url:http://www.aclweb.org/anthology/W/W09/W09-321.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, 11–21.

Sporns, O. (2002). Network analysis, complexity, and brain function. *Complexity, 8*(1), 56–60.

Sporns, O., Tononi, G., Edelman, G. M. (2002). Theoretical neuroanatomy and the connectivity of the cerebral cortex. *Behavioural Brain Research, 135*, 69–74.

Steyvers, M., & Tenenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 1*(29), 41–78.

Takamura, H., Inui, T., & Okumura, M. (2007). *Extracting semantic orientations of phrases from dictionary*. In C. L. Sidner, T. Schultz, M. Stone, & C. Zhai (Eds.), *HLT-NAACL* (pp. 292–299). The Association for Computational Linguistics.

Turtle, H. R., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transaction on Information System, 9*(3), 187–222.

Véronis, J., & Ide, N. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *COLING* (pp. 389–394).

Vitevitch, M. S., & Rodrguez, E. (2005). Neighborhood density effects in spoken word recognition in spanish. *Journal of Multilingual Communication Disorders, 3*, 64–73.

Voorhees, E. M., & Harman, D. K. (2005). *TREC: Experiment and evaluation in information retrieval*. MIT Press. url:http://trec.nist.gov/.

Wagner, A., & Fell, D. A. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London Series B Biological Sciences, 268*, 1803–1810.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications (structural analysis in the social sciences)*. New York: Cambridge University Press.

Watts, D., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440–442.

Widdows, D., & Dorow, B. (2002). *A graph model for unsupervised lexical acquisition*. In *COLING*.

Wilkinson, R., & Hingston, P. (1991). *Using the cosine measure in a neural network for document*. In Bookstein et al. (1991), pp. 202–210.

Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., et al. (2005) . Improving web search results using affinity graph. In Baeza-Yates et al. (2005), pp. 504–511.

Zhou, D., Schölkopf, B., & Hofmann, T. (2004). Semi supervised learning on directed graphs. In *NIPS*.