IR models based on network science: A research agenda

Giulio Leso and Stefano Mizzaro

University of Udine
Udine, Italy
giulioleso@gmail.com, mizzaro@uniud.it

Abstract. We propose to use network science concepts to build a novel IR model. We show some examples to explain the idea and to provide an intuitive rationale, and we sketch some future research directions.

1 Introduction

Information Retrieval (IR) models have been attracting a lot of research in the last 40 years: they are an important issue in textbooks, and they are still a hot issue in major IR conferences and journals. Network Science (NS) is a novel discipline, that has been steadily and quickly growing in the last decade. IR and NS are related by the well known PageRank and HITS algorithms (and similar ones like SALSA, etc.), that are exploited by search engines to rank the retrieved results, and are studied in both disciplines. These algorithms are instances of centrality measures, studied at length in NS and in Social Network Analysis; they can be better understood by the more general NS approach.

Our aim in this paper is to find a different kind of link between NS and IR: we try to devise a novel IR model based on NS concepts. More specifically, the aim is twofold: (i) to find that concepts that are well known in the IR models world, like, e.g., relevance, tf.idf, (pseudo-)relevance feedback, stopwords removal, have corresponding concepts in the NS world; and (ii) to show that these concepts in the NS world can be understood, analyzed, and extended in different — and hopefully more natural — ways than the corresponding ones in the IR world. If this happens, this analysis would of course allow a better understanding of basic IR concepts. This is a very preliminary work: no results are presented; we just try to provide some intuition that this research is sensible and we sketch our future work. We are not aware of any similar effort: the only slightly related work could be the attempts to use spreading activation in IR [4] and the application of the HITS algorithm to term stemming [2].

2 Background: Network science and IR models

NS [5,6] is an emerging and interdisciplinary research area that studies large-scale networks (or graphs) existing in the real world, like the Web, Internet, power grids, co-authorship, etc. These networks, because of their huge size, cannot be analyzed simply by drawing them; rather, more sophisticate statistical and algorithmic techniques are required to understand their properties, both of a static nature (e.g., which is the degree distribution of a network? Which is the average distance between two network nodes?) and of a dynamic one (e.g., what happens if the nodes represent living beings and the arcs represent contacts between them, and some nodes are infected with some disease: how will the epidemic spread? Will it affect the whole network?).

IR models have a long tradition, ranging from initial models (the boolean model, the probabilistic binary independence model, the vector space model) to more recent ones (BM25, language models, and quantum IR models). Over the years several important concepts have been added, like for example the importance of inverse document frequency, tf.idf, and document length normalization.

Another model, interesting for our purposes, has been described time ago in [7]. It is based on a neural network with three kinds of nodes: query-term (one node for each term in the query), vocabulary-term (one for each term in the collection), and document (one for each document in the collection). Query-term nodes are connected with directed edges to the same vocabulary-term, which are connected with directed edges to the document nodes that contain those terms. The connections activate first the vocabulary-term nodes, and then the potentially relevant documents. The weight of the edges can be calculated by applying tf.idf, to give higher weights to documents that contain more occurrences of a query term and to give less importance to terms occurring in many documents: this model is equivalent to the classical vector space.

The model can be immediately and naturally extended to include pseudo-relevance feedback by simply adding "inverse" directed edges that connect document nodes to vocabulary-term nodes (if the term is contained in the document). Thus, relevant documents may fire as well, activating other terms that will in turn activate other documents that may not contain query terms, but that may be relevant anyway because they contain the same terms that other relevant documents contain. The model can also be extended to include proper relevance feedback: more frequent terms are extracted from the documents judged relevant by the user, and added as query-terms; negative feedback can be modeled too, although it is more complex and requires removing the outgoing edges from the documents judged not relevant and re-running the query.

This model is a natural starting point to apply NS techniques, as the directed weighted network can be very large (billions of documents and millions of terms).

3 Relevance as a disease — and other ideas

One of the typical networks studied in NS features nodes that model individuals that are connected by an edge if they have physical contacts. In such a situation a contagious *disease* can spread over the network following the edges. The simplest formalization is the *SI model*, in which each node can be in one of two states: *Susceptible* and *Infected*. A susceptible node is a wealthy node which, if adjacent to an infected one, may become infected with a probability $\beta \geq 0$. Coming back to the neural network IR model, the relevance to the query can be modeled as an *epidemics* that runs through the network, starting from the query nodes and ending at the documents. According to the network connections, the disease will infect more or fewer nodes, giving rise to a more or less virulent epidemics.

By allowing β to vary according to the edge weights (tf.idf values), we can model a relevance disease that spreads more easily to those nodes that are connected by an edge having a high tf.idf value, i.e., those nodes, according to the IR model, that would give a high value of relevance to the documents. Pseudo-relevance feedback can also be modeled in the SI model by considering the "inverse" edges, from documents to terms, that allow the infected documents to pass on the disease to their term nodes, that will in turn infect other documents.

When considering vocabulary-term and document nodes and including the inverse edges, the neural network is *bipartite*. In NS, bipartite networks are often projected onto two *affiliation* networks: we can project our network onto two affiliation networks (one for terms and one for documents), which are interesting as well. In the former, representing term co-occurrence in documents, an epidemics can model automatic query expansion. In the latter, an epidemics can represent pseudo relevance feedback, or even proper relevance feedback if the user can label documents as relevant or not relevant. Since in relevance feedback a document initially judged relevant by the system can become not relevant after user input, the SI model seems not adequate, but its well known extensions seem useful. A first extension is the *SIS* model, that allows an infected node to go back, with a probability of $\gamma \geq 0$, to the initial state. Another extensions is the *SIR* model, where an infected node can become *Recovered*: this might model a not relevant judgment by the user. Vaccinations are also studied in this respect, and could better model negative relevance feedback.

In NS some results show that epidemics behavior depends on the shape of the network. For example, an epidemics is more likely to spread widely (and become pandemic) in a scale-free network than in a random network. Of course, it is not desirable that all nodes become infected/relevant: therefore, it would be interesting to study the shape of the above described networks. Although epidemics seems similar to spreading activation [4], the latter is more specific, and the former more embedded in general NS. Also, documents activated by pure spreading activation are query-independent [3] which is undesirable.

The relevance-as-a-disease metaphor looks interesting, but other NS methods can be used. For instance, the *information cascades*, that model the spreading of ideas, fads, etc., could be even more adequate, since information usually spreads not on the basis of individual contacts only, but in a more social way. For example, it is necessary that a certain fraction of the people connected to an individual adopt a fad to convince him to do so. The *cluster & cascade* theory could be useful in this respect: clusters of interconnected nodes are obstacles to cascades (i.e., relevance), since they behave as closed groups of individuals who may not be easily influenced by fashions. Also, besides relevance, other IR phenomena can be modeled by NS concepts; we briefly list some of them in the following.

Percolation models the removal of some nodes, and their incident edges, (*site percolation*) or of some edges (*bond percolation*). This model is useful to study phenomena like, e.g., breakdown of (or attack to) routers over the Internet, or temporary absence of communication channels between two cities. Nodes and edges can be removed at random, or according to specific criteria (like high degree nodes, heavy edges, central nodes). Site percolation on the vocabulary-term nodes can model stopwords removal: stopwords can be found automatically, using criteria like high degree nodes or central nodes, because they occur in most documents and therefore have a high number of in-edges. Also, this process would allow to remove all the non-discriminant terms, not the stopwords only, thus leading to a network that contains only meaningful words, which are good candidate for the index. This could be useful for index compression. Index compression might be obtained also by working on the affiliation network of terms and applying some clustering technique on it: this should be similar to LSI.

Another NS-IR link is between the recently proposed notion of retrievability [1], that studies how easy it is to retrieve a document in a collection, and *network search*, that, in NS, studies how to find a specific node in a network: retrievability could be modeled by search in the document affiliation network.

Finally, the issue of *ranking*, which is important in IR, for example to rank the retrieved documents or to rank the systems participating in a TREC-like evaluation competition, is studied extensively in NS: the links between ranking and voting are discussed in [5, Section 23.2].

4 Conclusions and future work

We have sketched how some (indeed, several) basic concepts and phenomena that are modeled in formal IR models can be seen with the NS glasses. These metaphors might be useful for: (i) a better understanding of the IR concepts; (ii) a generalization of these concepts, driven by NS methods; and (iii) a family of IR models based on network science, that will be then implemented and evaluated. Perhaps, we will also find some feedback from the IR world to the NS one.

References

1. L. Azzopardi and V. Vinay. Retrievability: an evaluation measure for higher order information access tasks. In *CIKM '08*, pages 561–570, New York, NY, USA, 2008.
2. M. Bacchin, N. Ferro, and M. Melucci. A probabilistic model for stemmer generation. *Information Processing and Management*, 41(1):121–137, Jan. 2005. Elsevier.
3. M. R. Berthold, U. Brandes, T. Kötter, M. Mader, U. Nagel, and K. Thiel. Pure spreading activation is pointless. In *CIKM '09*, pages 1915–1918. ACM, 2009.
4. F. Crestani. Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review*, pages 453–482, December 1997.
5. D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* University Press, Cambridge, July 2010.
6. M. Newman. *Networks: An Introduction.* Oxford University Press, USA, May 2010.
7. R. Wilkinson and P. Hingston. Using the cosine measure in a neural network for document retrieval. In *SIGIR '91*, pages 202–210, New York, NY, USA, 1991. ACM.