ABSTRACT

We propose a method for a concept-centric semantic analysis of an evolving corpus, highlighting the persistent concepts, emergence of new concepts, and the changes in the semantic associations between concepts. We report our findings on a corpus of computer science literature that spans six decades, revealing interesting patterns about the progress of the discipline.

General Terms

Algorithms, Languages.

Keywords

Distributional semantics, word embedding, natural language understanding, text mining, concept analytics.

1. INTRODUCTION

Recent progress in distributional semantics has led to several successful methods for obtaining vector representations of words in a language. The methods range from neural networks [1,6] to various matrix formulations [5,7]. Use of vectors has also been extended to represent larger pieces of text such as phrases, sentences, paragraphs and documents [6].

Our study carries these ideas further into the realm of concepts. Concepts are atomic units of meaning. A word sequence becomes a concept when it appears to encompass a significant meaning and presence in an area, in a historical context, in its relationships to other concepts, and in ways it influences perceptions. We explore sources for sequences of concepts in which the adjacency of concepts in the sequence is tied to closeness of the concepts in the semantic space. With these concept sequences, we use an embedding procedure to obtain vector representations for the concepts that capture their semantics. We call these semantic vectors for the concepts. In this paper we describe a way to use these semantic vectors to uncover facts about a corpus, and in particular, to trace the evolution of the corpus by observing the appearances of key concepts and changes in their semantic relationships. This is similar to the studies in [3,4] for temporal changes of word semantics, but our focus on concepts enables a richer characterization of the meaning and their changes.

2. CONCEPT SEMANTIC ANALYSIS

Syntactically, a concept is a single word or a word sequence (e.g., 'gravity', 'supreme court', 'Newton's second law'), which becomes a concept once it has been designated by a community to have a special role. A concept has many attributes: field of endeavor, origin, history, an associated body of work and/or knowledge, a cultural and/or historical connotation and more. So, although superficially, words, phrases and concepts seem similar, the latter embed a wider cultural context and a designation by a community. Concept sequences from text streams or exploratory behavior (e.g. click streams) allow for some normalization of diverse surface forms of an expression, and provide a richer characterization of the themes by the concept attributes.

One way of obtaining concept sequences is by applying an annotator (e.g., [2]) with a statistical decision procedure that is trained to recognize the chosen set of concepts as they appear in a text stream. The remaining words may be kept in place, resulting in a mixed words/concepts sequence, or left out, resulting in a concept-only sequence. In these sequences each recognized concept is replaced by a token identifying the concept. Sequences of both types can be sent into a standard word-embedding procedure to derive a vector for each concept or non-concept word. Embeddings derived from the full corpus, and those from the subsets at different snapshots, are expected to be different due to changes in the context surrounding each concept. Our analysis leverages these techniques to reveal the changes in semantic associations between the key concepts, so as to understand the essential evolution in the literature represented by this corpus. We develop the methods using a corpus in Computer Science due to our familiarity with the concepts therein; however, we believe that such analysis is applicable to many collections of scientific or creative literature, news reports, and historical archives of private documents.

3. DATA

The concepts used in this study are the titles of articles in the English Wikipedia, with redirections normalized to canonical forms. The vocabulary contains about 4.7 million concepts. An annotator for this set of concepts is trained with the links in the articles for each concept and their surrounding text.

The corpus is a collection of about 2 million abstracts of computer science related publications appearing in the years 1955-2014. We divided the corpus into 12 snapshots each covering 5 years, e.g., 1955-1959, 1960-1964, …, 2010-2014. The annotated sequences from the full corpus or from the snapshots are sent into an embedding procedure to derive a 200-dim semantic vector for each concept in the embedding space specific to the version of the corpus. Our analysis focuses on the concept associations within each space, represented by cosine similarity between the semantic vectors.

WWW 2016 Companion, April 11–15, 2016, Montréal, Québne Canada.
ACM 978-1-4503-4144-8/16/04.
DOI: http://dx.doi.org/10.1145/2872518.2889392.

4. PERSISTENT CONCEPTS

We first ask what key concepts have persisted through the years. We find 24 concepts common to every snapshot, and compute a hierarchical clustering of their semantic vectors derived from the full corpus. The dendogram is cut to produce 10 clusters. The most frequent concept in each cluster is selected as the cluster's representative. This results in 10 concepts: **Algorithm, Data, Mathematical model, Equation, Computer program, Solution, Computer, Language, Design, Mathematics.** The list displays a concentration in mathematical computation, due to the effect of enforcing the requirement that the concepts must have appeared in all temporal snapshots. Newer concepts about computers and computation appearing later are not part of this persistent set.

5. EMERGING CONCEPTS

To follow the emergence of important new concepts, we select the most frequently occurring concepts from each snapshot that are accountable for the top ¼ of all concept mentions. This results in 99 concepts that are rendered in Figure 1 by the first two principal components of their semantic vectors derived from the full corpus.

Figure 1 shows that many key concepts were laid down during the formative years of the discipline. The newest were brought in during 1995-2004: **Support vector machine, Web service, Wireless sensor network,** and **XML**. The newer concepts mostly occupy new regions in the semantic space (lower right corner) far away from the conventional areas in mathematics and programming. An exception is **Support vector machine,** which (colored orange) resides semantically among the older concepts in mathematics and statistics. Figure 1 also shows concentrations in mathematics and statistics, natural/programming languages, networking, etc., with no clean separation. This reflects the continuity between many of the apparently disjoint themes in the discipline. Further analysis on the manifolds formed by the semantic similarities may reveal more detailed structure.

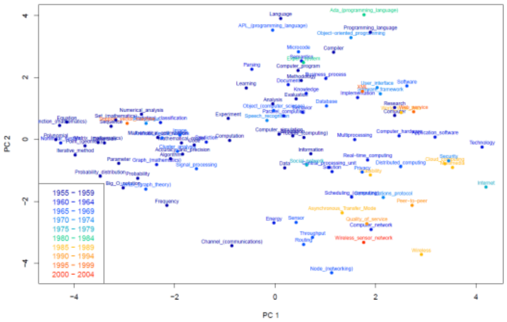

Figure 1: Key concepts colored by first appearance.

6. CHANGING SEMANTIC ASSOCIATION

The next question we ask is in what way concepts have changed in meaning over the years. While a formal definition of a concept is sometimes available in dictionaries (e.g. Wikipedia), its semantic content is often more fluid, and may change from time to time. We can follow such changes by observing the "semantic neighbors" of the concepts of concern in the per-snapshot embedding space. In Figure 2 we show changes (smoothed with a LOESS model) in association strengths of 30 concepts with the concept "**Computer network.**" The characteristic associations in each period are evident -- the rise of associations with Internet related concepts in the early 90's, followed by cloud computing, and later by mobile networks.

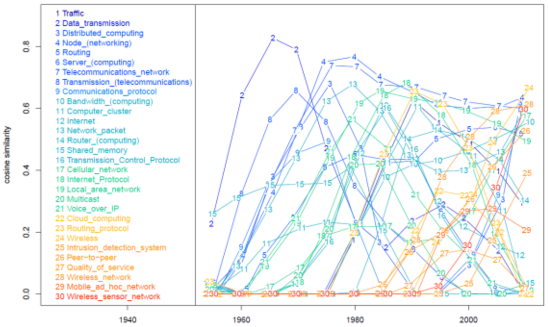

Figure 2: Changing associations with "Computer network."

7. CONCLUSIONS

We analyzed temporal changes in a corpus with a focus on the key concepts and their semantic associations. While we used the literature of computer science research as an example, we believe that the methodology is applicable to other time-stamped corpora.

8. ACKNOWLEDGMENTS

Our thanks to ACM, the curator of the collection of the abstracts, for sharing the corpus with us, and in particular Wayne Graves and Asad Ali for facilitating our access to the data. Our thanks also to Michele Franceschini and Livio Soares for their tools associated with the Concept Insights Service [8] of IBM Watson^{TM} for processing the corpus and the annotations.

9. REFERENCES

[1] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. 2003. A neural probabilistic language model. *J. Machine Learning Research.* 3, 1137-1155.

[2] Cheng, X. and Roth, D. 2013. Relational inference for Wikification. *Proc. of the Conf. on Empirical Methods in Natural Language Processing* (Seattle, Washington, USA, October 18–21, 2013), 1787-1796.

[3] Jatowt, A. and Duh, K. 2014. A framework for analyzing semantic change of words across time. *Proc. of the ACM/IEEE-CS Joint Conf. on Digital Libraries* (London, UK, September 8-12, 2014), 229-238.

[4] Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. 2015. Statistically significant detection of linguistic change, *Proc. of the Int'l World Wide Web Conf.* (Florence, Italy, May 18-22, 2015) , 625-635.

[5] Levy, O. and Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. *Proc. of the Annual Conf. on Neural Information Processing Systems* (Montreal, Quebec, Canada, December 8-13, 2014), 2177-2185.

[6] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Proc. of the Annual Conf. on Neural Information Processing Systems* (Lake Tahoe, Nevada, USA, December 5-10, 2013), 3111-3119.

[7] Pennington, J., Socher, R., and Manning, C.D. 2014. Glove: Global vectors for word representation. *Proc. of the Conf. on Empirical Methods in Natural Language Processing* (Doha, Qatar, October 25–29, 2014), 1532-1543.

[8] http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/concept-insights.html.