Extracting the Dynamic Popularity of Concepts from a Corpus of Short-Sentence Documents

Willy Picard

Extracting the Dynamic Popularity of Concepts from a Corpus of Short-Sentence Documents

Willy Picard

Department of Information Technology,
Poznań University of Economics,
al. Niepodległości 10,
61-875, Poznań, Poland
picard@kti.ue.poznan.pl

Abstract. The decomposition of information into smaller bunches of data is a commonly observed process on the Web, Twitter and RSS being manifestations of this process. As a consequence, a shift may be observed from an information world in which information comes in large bunches of data, to a world of short-sentence documents. This shrinking of information chunks goes along with an explosion of the number of these chunks. Therefore, information may often be aggregated in corpuses of documents consisting of many short sentences. The identification of important concepts in corpuses of short-sentence documents is a difficult, but necessary, task to understand the whole information. Understanding the dynamics of the popularity of important concepts is necessary to capture the evolution of the corpus in time. In this paper, a method to extract the important concepts from a corpus of short-sentence documents is proposed. A model of the popularity of concepts and its dynamics is proposed, together with an algorithm to analyze the dynamics of important concepts. Finally, the proposed method is validated with an analysis of the titles of the articles published at eleven IFIP Working Conferences on Virtual Enterprises, from PROVE'99 to PROVE'10.

Keywords: text mining, context extraction, collaborative network, virtual enterprise, virtual organization, dynamic popularity

1 Introduction

A major shift in the way information is designed, produced, sold, and consumed may currently be observed. Information is currently decomposed in smaller bits of data. Instead of newspapers, single articles are written, published, sold, and read. Instead of CDs containing a list of songs, music is produced, sold, and composed in form of individual songs, as MP3 files.

The decomposition of information goes together with the production of very short bunches of data. Some websites, such as Twitter, enforce the production of small bunches of data. Twitter [1] limits the length of messages, referred to as "tweets", to 140 characters. Similarly, all major social networking websites, such as Facebook [2] and Google+ [3], provide their users with the possibility to provide a short

information concerning their "status". Additionally, the graphical user interface used to enter data or posts is usually limited to a small input text field. As a consequence, most posts submitted to social networking websites are short.

Finally, the trend towards shorter, decomposed information goes further with the mechanism of information summarizing. Technologies such as RSS [4] and Atom [5, 6] provide a means to summarize information, usually to a few dozens or hundreds words. Although the original purpose of these technologies was the possibility to annotate websites, providing semantic meta-data for a further computer processing, RSS and Atom feeds are currently used mainly to syndicate and aggregate information for humans, especially with the rise of mobile computing.

Structuring information in small bunches of data goes together with a drastic rise of the number of bunches of data associated with a given topic. As a consequence, information is organized as set of very numerous and short bunches of data, often consisting of single sentences. In this paper, such sets of bunches of data are referred to as Corpus of Short-Sentence Documents.

> A *Corpus of Short-Sentence Documents (CSSD)* is defined as a time-indexed list of sets of documents, with each document limited to a high number of short sentences.

Examples of CSSDs may be the results of a Twitter search on a given topic, the list of email subjects in a given folder, and a list of lecture subjects offered by a university grouped by years.

The decomposition of information in CSSD leads to important challenges for their consumers. A first challenge is the identification of key concepts in the CSSD. In a world of not-decomposed information, the key concepts are explicit in the structure itself: the titles of chapters in books are usually focusing on the key concepts presented in the contents of the chapters. In newspapers, various sections and the titles of the articles emphasize the key concepts. In CSSD, no structural entity is available: no title or sections are presented. Therefore, the identification of key concepts requires the whole corpus to be analyzed, which is challenging because of the number of documents it contains.

A second challenge is the understanding of the dynamics of the popularity of concepts in the CSSD. CSSDs should be considered as streams, with new documents continuously enriching the corpus. Therefore, the popularity of a concept usually evolves in time, as new documents are added to the corpus. The popularity of concepts is a dynamic variable, having various values in time.

In this paper, a method to extract the important concepts from a corpus of short-sentence documents is proposed. A model of the popularity of concepts and its dynamics is proposed, together with an algorithm to analyze the dynamics of important concepts.

The rest of this paper is organized as follows: in Section 2, the concepts of CSSD and popularity are defined, followed by the presentation of our research goal. In Section 3,

the proposed method is presented, In Section 4, the proposed method is validated with an analysis of the titles of the articles published at eleven IFIP Working Conferences on Virtual Enterprises, from PROVE'99 to PROVE'10 [7-17]. Finally, Section 6 concludes the paper.

2 Research Goal

2.1 Fundamental Definitions

The presentation of our research goal requires a precise definition of CSSDs and dynamic popularity.

A *CSSD*, further denoted as γ, is a list of time-indexed documents, i.e.,

$$\gamma = \{d_t\},$$

where d_t is a document indexed by time t.

A *time-indexed document* d_t consists of a set of sentences S_t and a time index t, i.e.,

$$d_t = < S_t=\{s_{t,n}\}, t >,$$

where $s_{t,n}$ is the n-th sentence of the document indexed by time index t.

A *sentence* $s_{t,n}$ is a limited list of characters. The maximal number of characters of sentences depends on the type of CSSD. For instance, in CSSDs containing RSS 0.91 item titles (resp. item descriptions), the maximal number of characters of sentences is 100 (resp. 500) characters. In CSSDs containing Twitter "tweets", the maximal number of characters of sentences is 140 characters.

A *concept* c is defined as a non-stop word stem. Stop words are most common words, such as "and", "the", and "for" in English. Stems are the base of inflected and derived words. For instance, the words "cooperate", "cooperation", and "cooperating" share the same stem "cooper".

The *static popularity* of a concept c in time-indexed document d_t, further denoted as $p_{c,t}$, is defined as the index of the concept c in the popularity ranking of d_t. The popularity ranking of d_t, further denoted as $p_{d,t}$, is the list of concepts of d_t ordered by the number of their occurrences.

The *dynamic popularity* of a concept c in the corpus γ, further denoted as π_c, is defined as a vector containing the indexes of the concept c in the popularity global rankings of d_t ordered by time. The popularity global ranking of d_t, further denoted as $\pi_{d,t}$, is the list of concepts of γ ordered by the number of their occurrences in d_t.

2.2 Research Goal

Our research goal is to develop a method to extract the dynamic popularity of concepts from a CSSD. The considered CSSD are monolingual, i.e., all the sentences of all the time-indexed documents are written in the same language. The method should be independent of the language of the CSSD. The method should be fully automatic and should not require any human action. An appropriate graphical representation should provide a means for a better understanding of the results of the method.

3 A Method to Extract Dynamic Popularity of Concepts in a CSSD

The proposed method to extract dynamic popularity of concepts in a CSSD consists of three steps: data preparation, extraction of popular concepts, and extraction of dynamic popularity.

3.1 Data Preparation

The first step of the proposed method aims at preparing the data for further text mining. It is assumed that a CSSD has been formerly gathered and compiled in an appropriate digital form. The preparation phase starts by the lower case conversion of all the sentences of all the documents in the CSSD. Next, white space is removed, together with punctuation marks. Then, stop words are removed, based on a formerly prepared list of stop words for the language of the CSSD. Different lists of stop words have to be used to prepare CSSDs written in different languages. Finally, all the sentences are stemmed, i.e., all lowercase, whitespace-free non-stop words are replaced by their stems. The widely used algorithm for stemming proposed by Porter [18] is suggested as a method for the stemming of CSSD, but any other stemming algorithm may be integrated to the method.

The result of the first step of the method is a cleaned concept corpus γ', that consists of cleaned time-indexed documents containing cleaned sentences. A *cleaned sentence s'* is a list of concepts {c'}.

3.2 Extraction of Popular Concepts

The next step aims at identifying the most popular concepts. The extraction of popular concepts is proposed as a bottom-up process, i.e. popular concepts are first identified for each time-indexed document, and next, all the identified popular concepts are merged into one common set of popular concepts.

The identification of popular concepts for a given time-indexed document d_t consists in selecting the first elements of the popularity ranking of d_t. A term-document matrix of the CSSD is computed to establish the popular ranking of d_t. A term-document matrix *tdm* is a matrix whose values are the number of occurrences of a given concept (given in columns) in a given document (given in rows), i.e.,

$$tdm_{t,c} = \sum_{s' \in S'_t} |\{c' \in s' : c' = c\}|$$

The number of occurrences of concepts in a given time-indexed document are given is the associated row of the term-document matrix. The popularity ranking of d_t is therefore the values of the sorted row associated with d_t in the term-document matrix. The establishment of the set of most popular concepts is based on the popularity rankings for all the time-indexed documents: the most popular concepts of each popularity ranking are merged together to create the set of most popular concepts. An important parameter of the method is the number of popular concepts to be kept from each popularity ranking in the set of the most popular concepts. This parameter is further denoted α.

Formally, the set of most popular concepts C_{pop} for a given value of α is such that,

$$c \in C_{pop} \Leftrightarrow \exists t : tdm_{t,c} \geq \alpha.$$

3.3 Extraction of Dynamic Popularity

The extraction of dynamic popularity is based on the processing of the *popularity matrix* from the term-document matrix. The popularity matrix *pm* is a matrix whose values are the ranking of a given popular concept (given in columns) in a given document (given in rows). The ranking of a popular concept *c* in a document d_t is the index of the concept in the sorted row associated with d_t in the term-document matrix. Therefore the most popular concept of a given document, i.e., the concept that has the larger number of occurrences in this document, has a ranking equals to 1. The second most popular concept has a ranking equals to 2, etc. Therefore, each column of the popularity matrix contains the ranking of concepts in a given document, while each row of the popularity matrix contains the various ranking values of a given concept across documents. Rows of the popularity matrix are dynamic popularity of the associated concept.

3.4 Summarizing the Proposed Method in Pseudo-code

The proposed solution may be summarized in pseudo-code as follows:

```
corpus <- Read(corpusSource)

lowercase(corpus)
removeWhiteSpaces(corpus)
removePunctuation(corpus)
removeStopWords(corpus)
stem(corpus)

tdm <- processTermDocumentMatrix(corpus)
popularConcepts <- emptySet()
```

```
foreach row in tdm
  sort(row)
  foreach concept in row
    if (tdm(row,concept) ≥ α)
      popularConcepts.add(concept)

pm <- emptyMatrix()
foreach row in tdm
  sort(row)
  foreach concept in popularConcepts
    pm(row,concept) = row.indexOf(concept)
```

The dynamic popularity of a given concept is the row of the matrix *pm* associated with this concept.

4 Validation of the Proposed Method

The proposed solution has been applied to a corpus containing the titles of the articles published in the proceedings of the eleventh first editions of the IFIP Working Conferences on Virtual Enterprises, from PROVE'99 to PROVE'10 [7-17]. The PROVE CSSD contains 721 articles, with an average number of 65.6 articles per conference edition. The CSSD contains 6620 words. The proposed method has been implemented with the *R* software environment for statistical computing and graphics [19]. The associated package *tm* [20] provides support for most required functions, such as stop words removal, stemming, term-document matrix processing.

After the preparation step, the set of concepts is reduced to 1031. Next, the 20 most popular concepts for each edition have been identified, leading to a set of 68 popular concepts. The 5 most popular concepts for PROVE'99, PROVE'05, and PROVE'10 are presented in Table 1.

Table 1. Five most popular concepts in articles published in the proceedings of PROVE'99, PROVE'05, and PROVE'10.

Popularity	PROVE'99	PROVE'05	PROVE'10
1	enterpris	Virtual	collabor
2	virtual	Collabor	network
3	manag	Network	service
4	prodnet	Organ	support
5	infrastructure	Model	system

Next, the dynamic popularity of the 68 identified popular concepts has been processed. The dynamic popularity of the chosen concepts "servic", "collabor", "network", "approach", "infrastructur" is presented in Table 2.

Table 2. Dynamic popularity of five popular concepts.

Concept	’99	’00	’02	‘03	‘04	‘05	‘06	‘07	‘08	‘09	‘10
service	20	20	20	20	20	10	10	9	9	10	3
collabor	20	20	15	11	3	2	2	1	1	1	1
network	20	8	9	9	2	3	1	3	2	2	2
approach	20	20	20	6	11	6	15	20	6	8	7
infrastructur	5	5	6	20	20	20	20	20	20	20	20

The dynamic popularity of these concepts is presented graphically in Figure 1.

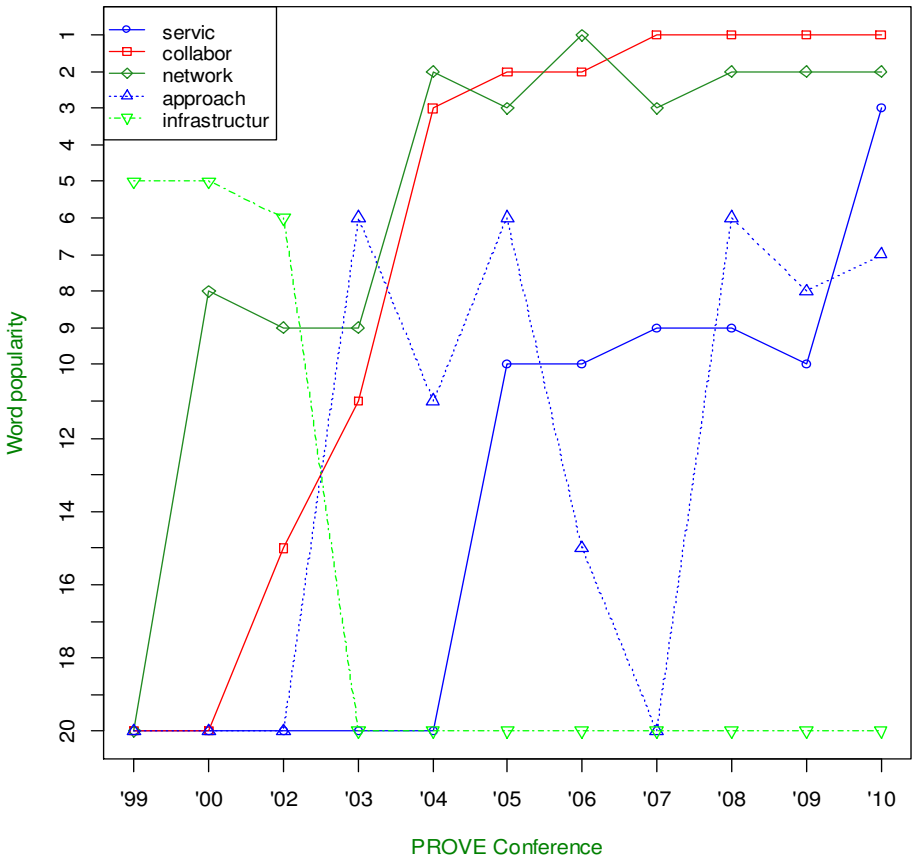

Fig. 1. Dynamic popularity of five popular concepts.

A map of all the identified popular concepts is presented in Figure 2. Concepts are located on the x axis according to the difference between their popularity in PROVE’10 and PROVE’99. On the y axis, concepts are plotted according to the variance of the differences of their popularity between two consecutive PROVE editions. The emerging concepts, whose emergence is stable are on the top-right quadrant, e.g., concepts "servic", "collabor", "network". Extinguishing (in some case

extinguished) concepts are on the left side of the figure, e.g., concept “infrastructure”. The dynamics of concepts represented in the lower part of the figure is turbulent, e.g., the dynamics of the concepts “approach”.

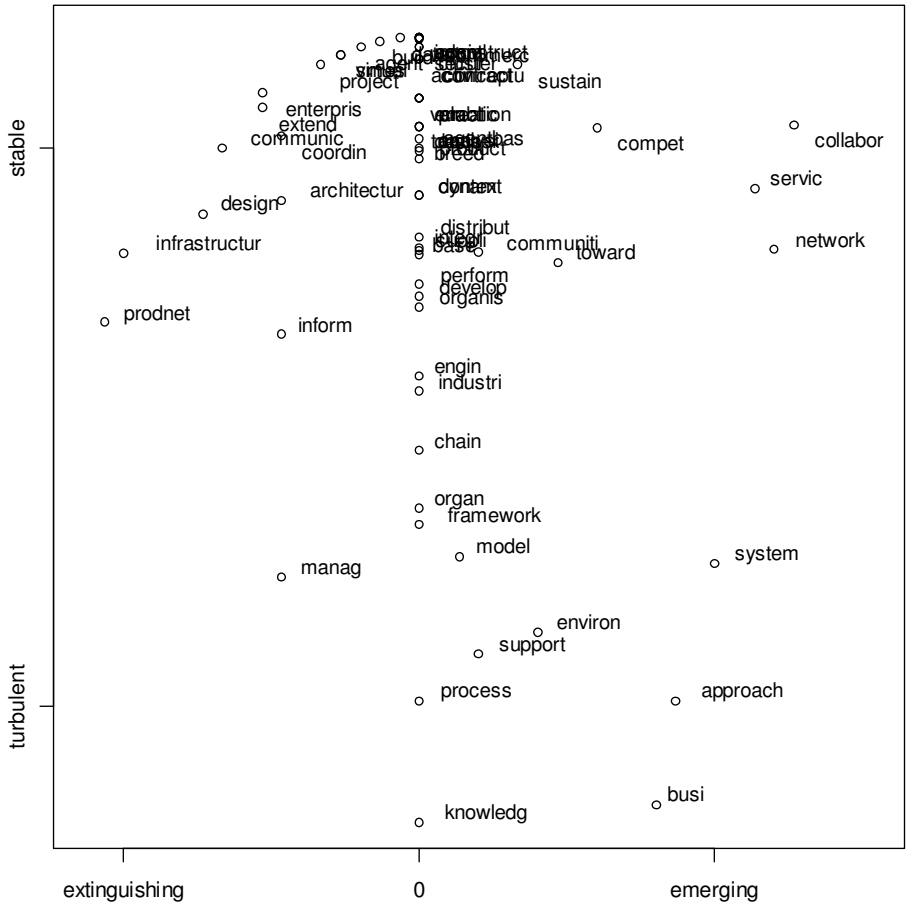

Fig. 2. A map of concepts according to their popularity.

The map of concepts presented in Figure 2 illustrated the shift in core concepts used in the community attending PROVE conferences, from “virtual” “enterpris”, to “collabor” “network”. An additional remark concerns the identified important increase of the popularity of the concept “service”, confirming the pertinence of the main topics of the PROVE’12 conference.

5 Conclusion

In this paper, a method is proposed for the extraction of dynamic popularity of concepts in CSSDs. It has been demonstrated that the application of the proposed method to the CSSD consisting of the titles of the papers published in the proceedings of the consecutive editions of the PROVE conferences leads to the identification of trends concerning the concepts used by the community attending these conferences. The shift from "virtual enterprises" to "collaborative networks" appears clearly in the results of the proposed method.

The proposed method may be applied not only to the title of other conference series, but also to other CSSDs. An example may be the identification of trends in the results of Twitter searches, or RSS channels.

It should also be noted that the proposed method is independent of the language of the given CSSD. The only requirement for the method to support a given language is the existence of a list of associated stop words and appropriated stemming algorithms.

A main limitation of the proposed method is its limitation to single-term concepts, e.g., "cooper". The extension to multi-term concepts, e.g., "cooperative network" is an area that should be further studied.

In future works, dynamic popularity should be normalize as regards the number of sentences or the number of concepts in a given document. Currently only the number of occurrences of a given concept in a given document is taken for the dynamic popularity.

Finally, it would be interesting to consider additional information, such as the abstract or the keywords in the identification of the dynamic popularity of concepts in a CSSD.

Acknowledgments. The author wishes to thank Luis Camarinha-Matos for providing the list of papers published in the proceedings of the eleven PROVE editions and studied with the proposed method.

References

1. Twitter, http://www.twitter.com/
2. Facebook, http://www.facebook.com/
3. Google+, http://plus.google.com/
4. Libby, D.: RSS 0.91 Spec, revision 3, Netscape Communications, July 10, 1999, http://www.rssboard.org/rss-0-9-1-netscape
5. Internet Engineering Task Force (IETF): The Atom Syndication Format, RFC 4287. Nottingham, M., Sayre, R. (eds.) http://tools.ietf.org/html/rfc4287 (2005)
6. Internet Engineering Task Force (IETF): The Atom Publication Protocol, RFC 5023. Gregorio, J., de hOra, B. (eds.) http://tools.ietf.org/html/rfc5023 (2007)
7. Camarinha-Matos, L. M., Afsarmanesh, H. (eds.): PRO-VE '99: Proceedings of the IFIP TC5 WG5.3 / PRODNET Working Conference on Infrastructures for Virtual Enterprises: Networking Industrial Enterprises. Kluwer, B.V. Deventer, The Netherlands (1999)

8. Camarinha-Matos, L. M., Afsarmanesh, H., Rabelo, R. (eds.): PRO-VE '00: Proceedings of the IFIP TC5/WG5.3 Second IFIP Working Conference on Infrastructures for Virtual Organizations: Managing Cooperation in Virtual Organizations and Electronic Business towards Smart Organizations: E-Business and Virtual Enterprises: Managing Business-to-Business Cooperation: Networking Industrial Enterprises. Kluwer, B.V. Deventer, The Netherlands (2000)
9. Camarinha-Matos, L. M. (eds.): PRO-VE '02: Proceedings of the IFIP TC5/WG5.5 Third Working Conference on Infrastructures for Virtual Enterprises: Collaborative Business Ecosystems and Virtual Enterprises. Kluwer, B.V. Deventer, The Netherlands (2002)
10. Camarinha-Matos, L. M., Afsarmanesh, H. (eds.): PROVE'03: Processes and Foundations for Virtual Organizations, IFIP TC5/WG5.5 Fourth Working Conference on Virtual Enterprises, PRO-VE'03. Kluwer (2003)
11. Camarinha-Matos, L. M. (eds.): Virtual Enterprises and Collaborative Networks, IFIP 18th World Computer Congress: IFIP TC5/WG5.5 5th Working Conference on Virtual Enterprises, PRO-VE'04. Kluwer (2004)
12. Camarinha-Matos, L. M., Afsarmanesh, H., Ortiz, A. (eds.): Collaborative Networks and Their Breeding Environments: IFIP TC 5 WG 5. 5 Sixth IFIP Working Conference on VIRTUAL ENTERPRISES, PRO-VE'05. Springer-Verlag (2005)
13. Camarinha-Matos, L. M., Afsarmanesh, H., Ollus, M. (eds.): Network-Centric Collaboration and Supporting Frameworks: IFIP TC 5 WG 5.5, Seventh IFIP Working Conference on Virtual Enterprises, PRO-VE'06. Springer-Verlag (2006)
14. Camarinha-Matos, L. M., Afsarmanesh, H., Novais, P., Analide, C. (eds.): Establishing the Foundation of Collaborative Networks: IFIP TC 5 Working Group 5.5 Eighth IFIP Working Conference on Virtual Enterprises, PRO-VE'07. Springer (2007)
15. Camarinha-Matos, L. M., Picard, W. (eds.): Pervasive Collaborative Networks: IFIP TC 5 WG 5.5 Ninth Working Conference on VIRTUAL ENTERPRISES, PRO-VE'08. Springer (2008)
16. Camarinha-Matos, L. M., Paraskakis, I., Afsarmanesh, H. (eds.): Leveraging Knowledge for Innovation in Collaborative Networks: 10th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2009. Springer (2009)
17. Camarinha-Matos, L. M., Boucher, X., Afsarmanesh, H. (eds.): Collaborative Networks for a Sustainable World - 11th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2010. Springer (2010).
18. Porter, M.F.: An algorithm for suffix stripping. Program. 14, 130--137 (1980)
19. The R Project for Statistical Computing, http://www.r-project.org/
20. CRAN - Package tm, http://cran.r-project.org/web/packages/tm/