A Modified Tripartite Model for Document Representation in Internet Sociology

Mikhail Alexandrov[1,2], Vera Danilova[1,2(✉)], and Xavier Blanco[1]

[1] Autonomous University of Barcelona, Barcelona, Spain
{MAlexandrov.UAB,XblancoE}@gmail.com
[2] Russian Presidential Academy of National Economy and Public Administration, Moscow, Russian Federation
vera.danilova@e-campus.uab.cat

Abstract. Seven years ago Peter Mika (Yahoo! Research) proposed a tripartite model of actors, concepts and instances for document representation in the study of social networks. We propose a modified model, where instead of document authors we consider textual mentions of persons and institutions as actors. This representation proves to be more appropriate for the solution of a range of Internet Sociology tasks. In the paper we describe experiments with the modified model and provide some background on the tools that can be used to build it. The model is tested on the experimental corpora of Russian news (educational domain). The research reflects the pilot study findings.

Keywords: Tripartite model · Document representation · Ontology · Social networks analysis · Internet sociology

1 Introduction

The automatic detection of document and corpus topic is one of the most popular tasks within the natural language processing. The traditional model, where each document is represented by a vector describing the distribution of keywords in a document or corpus, works quite well. On the basis of this representation, it becomes possible to group similar documents and reveal the corpus structure, as well as to group keywords and build topic vocabularies [6,11]. Considering each document as an ontology instance and a set of keywords as a concept, we obtain a two-layer ontology model. The appearance of social networks allowed the researchers to bring in the social aspect, thus creating a tripartite model of ontology for Social Networks Analysis containing document authors (actors), document content (instances), and tags (concepts) [13]. Three-layer model is a tool for studying the emergence and dynamics of communities on the basis of user-generated content. Author profiling gave an opportunity to analyze the

Under partial support of the Catholic University of San Pablo (grant FINCyT-PERU).

Q. Chen et al. (Eds.): DEXA 2015, Part II, LNCS 9262, pp. 323–330, 2015.
DOI: 10.1007/978-3-319-22852-5_27

attitude of different user categories towards particular events and topics. However, it is not only the author's personality that matters for solving Internet Sociology tasks, but also the awareness of the connection between the text mentions of personalities and organizations and concepts and instances within text. This knowledge is important, when dealing with traditional mass media - e-publications created by analysts and journalists. In this case, the Actor in the Actor-Concept-Instance triple is a Named Entity (NE) mention (person and organization names), the Concept is an automatically generated topic description, not user-generated tags, and the Instance is a set of phrases with a high level of specificity that frequently denote events (olympiad, demonstration, etc.). The present paper considers the use of this modified tripartite model of ontologies for the purposes of Internet Sociology studies.

The rest of the article is organized as follows. Section 2 provides a description of the three-layer model. Section 3 presents the experiments on real-world data that include model construction, opinion mining and clustering. Section 4 concludes the paper. This paper outlines our pilot-study experiments performed on the basis of Russian newspaper articles. The listed examples are translated into English.

2 Multipartite Models

2.1 Tripartite Model

The tripartite model uses three layers of descriptors (keywords) for document representation as follows. The first layer is formed from a list of actors or person and organization names that are mentioned in a document; the second layer is a list of concepts that characterize the domain a given document belongs to; the third layer is represented by the instances or document-specific collocations. Each element of each list contains a set of tokens, which number varies from 1 to 5. As for a single document, the tripartite model can be created for a corpus of documents. The corpus representation is a list of actors, concepts and instances that are mentioned throughout a corpus. Table 1 presents an excerpt from a model for a corpus of 250 documents related to the educational domain.

Here, the NE "Dr. Livanov" is a mention of the Russian Minister of Education and Science. "Dr. Rukshin" is one of the leading Russian school teachers. Parliament means a discussion at the State Duma. President means here a discussion at the Public Council under the President's administration. Moscow means the Department of Education of the Moscow Regional Government. It should be

Table 1. An excerpt from our experimental corpus model (educational domain)

No.	Actor	Concept	Instance
1	*Ministry of Education*	*Young Talent*	*Law of Education*, Parliament, Mar. 2013
2	*Cabinet of Ministers*	*Primary School*	*Law of Education*, President, Jul. 2014
3	*Higher School of Economics*	*Secondary School*	*College Education*, Parliament, Dec. 2013

noted that the frequency of occurrence of key terms in a document is not taken into account when building the model. We focus on the presence or absence in a document of the keywords from the corpus-based model.

2.2 Bipartite Model

The notation for the dimensionality of each layer (list) for the corpus of documents is as follows. Let K_a be the number of actors, K_i - the number of instances, and K_c - the number of concepts.

Actors-Concepts model (AC model) shows what keywords each person or organization is connected with in a text. It can be presented as a table (matrix) of size M (K_a, K_c). Matrix transposition MT of size (K_c, K_a) will show the actors that are related to a given concept (CA model).

The connection between an actor and a concept can be evaluated using the Jaccard distance [6]:

$$J(A_i, C_j) = 2 \times \frac{|D_{A_i} \cap D_{C_j}|}{|D_{A_i} \cup D_{C_j}|}, \tag{1}$$

where A_i and C_j represent a specific actor and a concept respectively, D_{A_i} and D_{C_j} correspond to the collections of documents, where the mentioned concept and actor occur. The number of documents, where the given concept and actor co-occur, in the numerator is doubled in order to normalize the connection degree from [0.0, 1.0].

A weight W_A for each actor on the concepts layer, and a weight W_C for each concept on the actors layer can be introduced as follows: $W_A = \sum_{i=1}^{K_{cj}} J(A_i, C_j)$ (K_{cj} is the number of concepts related to a specific actor, and $J(A_i, C_j)$ is the Jaccard distance between the selected actor and one of the related concepts) and $W_C = \sum_{i=1}^{K_{aj}} J(C_i, A_j)$ (K_{aj} is the number of actors related to a specific concept, and $J(C_i, A_j)$ is the Jaccard distance between the selected concept and one of the related actors).

Concepts-Instances model (CI model) shows the specific collocations each concept (keyword set) is related to and vice versa. The weights are introduced in the same way as for the previous models. CI and IC models are commonly used for natural language processing purposes in two-layer model-based document parametrization: CI matrix shows the distribution of keywords throughout the documents, and IC - specific expressions that are "spotted" in each of the documents.The *Instances-Actors* (IA) and *Actors-Instances* (AI) models can be built by transforming the previously obtained matrices: $(AI) = (AC) \times (CI)$ and $(IA) = (IC) \times (CA)$.

2.3 Unipartite Model

Two-layer models can be reduced to single-layer models, which further allows to perform automatic grouping. A one-layer model uses only one of the layers

or dimensions (actors, concepts or instances) for document and corpus representation. In order to properly build the unipartite model, let us first consider the (AC) model, where we can find the connections between two actors on the concept layer. To this end, the calculation of the number of shared concepts for each pair of actors should be performed. If all of the concepts are the same, the connection will have its maximum weight, and if no concepts are shared, the connection is insignificant. The connection can be evaluated using the Jaccard distance as follows:

$$J(A_i, A_j) = 2 \times \frac{|K_{ci} \cap K_{cj}|}{|K_{ci} \cup K_{cj}|}, \tag{2}$$

where A_i and A_j are the corresponding actors, K_{ci} and K_{cj} are sets of concepts related to the actors A_i and A_j correspondingly. The number of shared concepts in the numerator is doubled in order to normalize the connection degree from [0.0, 1.0].

Having performed a pairwise evaluation of actors connection on the concepts layer, we can build a matrix Actors - Actors (AA), which is the target unipartite model. The resulting proximity matrix is the input for the cluster analysis. As a result, we obtain the groups of interconnected actors. It is a meaningful result: each cluster contains person and organization names discussed within the same topics. In a similar way, the reversed (CA) matrix provides the Concepts-Concepts model (CC). In this case, clustering also produces a meaningful result: the connections between the concepts that share the same actors can be seen. In a similar fashion, we obtain the matrices Concepts-Concepts (CC) and Instances-Instances (II) from the bipartite (CI) model, and Actors-Actors (AA) and Instances-Instances (II) from the bipartite Concepts-Instances model. As it can be seen, (AA), (CC) and (II) matrices have been obtained twice. The difference is that (AA) in one case reflects the connection between the actors on the concept layer, and in the other case - on the instance layer, and similarly for the matrices (CC) and (II).

3 Experiments

A corpus of 250 articles on education has been compiled for testing. We consider this amount enough for a pilot study, because we can easily check the performance of clustering and opinion measurement algorithms. An excerpt from an article on education (translated) is given in Fig. 1. In this text we can find the elements belonging to the three layers: S. Rukshin, A. Kiryanov, D. Medvedyev (Actors); Education (Concepts); gifted children, lyceums (Instances).

We consider two main operations on the obtained models: clustering (unipartite model) and opinion polarity analysis (tripartite and bipartite models).

3.1 Model Construction

Document models are built based on a corpus model, which includes 3 keyword lists forming the actor, concept and instance dimensions. The corpus is created

LYCEUM NETWORKS

In 2009, a famous Russian teacher S. Rukshin from Saint-Petersburg and his colleague A. Kiryanov developed a strategy for the education of gifted children and presented it to the President D. Medvedyev. Firstly, they proposed to create a network of federal lyceums for generally gifted children. Secondly, for the children with specific talents in mathematics, physics, etc., they suggested opening a network of lyceums under the biggest universities of the country ...

Fig. 1. A sample document (educational domain)

using keyword-based queries, where the keywords cannot be randomly selected and should represent a description of a certain topic. The three-layer model construction includes two steps: (i) identify the descriptors (keywords and keyphases denoting actors, concepts and instances within the given corpus) for each layer; (ii) parametrize texts in the three dimensions.

Identification of Descriptors

The extraction of actors and instances is a named entity recognition (NER) task. As a solution, we extract single terms on the basis of their specificity criterium using LexisTerm-I [2]. Term frequencies can be counted either in "Corpus" mode for the whole corpus or in "Document" mode for each particular document. In this implementation, LexisTerm-I proposes some candidates for the construction of actor, instance and concept lists, and an expert does the manual correction of the obtained list. At a later stage, we plan to make the procedure totally automatic using approaches described in [8–10].

Actors List. The frequency of actors in the domain-specific documents is assumed to be higher than in the general lexis, therefore, they are detected in the "`Document`" mode. Our education-related corpus contains $n = 50$ actor names, where there are 30 person names and 20 organization names. Each list element includes 1–5 tokens, which allows to record full names with titles (e.g., Dr. Dmitry Livanov), and full organization names (e.g., Ministry of Science and Education).

Concepts List. The terms have been extracted by LexisTerm-I for $K = 10$ in the "`Corpus`" mode. Each concept in the resulting list includes one or several words. The sample main concepts are as follows: education accessibility, young talents, Unified State Exam, etc.. Within the "Education" domain we distinguish $M = 10$ subtopics, for which keyword lists are manually created. The number of descriptors cannot be higher than 20, which, from our experience, is enough. The number of tokens per concept together with descriptors varies from 1 to 5.

Instances List. The instance list is constructed using the results of LexisTerm-I text processing in the "`Document`" mode. Each instance is represented by 3 components: (i) an object or event; (ii) object/event location; (iii) object/event date. The sample instances are the Law of Education, the decree on the creation of a lyceum network, Moscow school merging, etc. There is also a fixed list of locations: Parliament, President's Council, Moscow Government, etc. The date

includes the related month and year. The resulting list includes thus the following components: $[object_1, object_2, ...object_n]$, $[location_1, location_2, ...location_n]$, and *(month, year)*. For our corpus, the list of objects and events contains $P = 20$ elements, the list of locations - $Q = 15$ elements. Each element's length is of 1–5 tokens.

Document Parametrization

Document parametrization is the representation of a document in the three dimensions of actors, concepts, and instances on the basis of the corpus model performed with the *ParamDoc-3D* program. Operations on two dimensions are performed in *ParamDoc-2D*, and on one dimension - in *ParamDoc-1D* [1]. The base tripartite model is parametrized as follows.

Actor Dimension. All actor mentions from the N-actor list of the corpus model in a document are collected. The actor mention is selected if it corresponds to the expert-established settings. For unigrams, a complete match is obviously required. In case of 2–5 token keywords, a complete or partial match (2-3 tokens) can be set.

Concept Dimension. The weight of each topic in a document is measured using a vocabulary of the corresponding descriptors. A topic with the highest weight, or topics with the weights exceeding certain threshold are selected. The calculation uses the following method [5]: (a) measure the text coverage by the vocabulary (total density of the descriptors), (b) measure the vocabulary coverage by the text (percentage of vocabulary entries' mentions), (c) measure the total coverage. Both coverages can take on a value $[0.0, 5.1]$ using a non-linear scale. The total value can range from 0 to 2. By establishing a threshold, we can sort out a set of topics present in a document.

Instance Dimension. Instance extraction is done in 3 steps: (1) search for an event or object from the list of P-instances of the corpus model, (2) search for location triggers from the list of Q-locations of the corpus model, (3) search for the mentions of month and year in the document body. Objects, events, and locations are extracted similarly to actors: total and partial matches of the 1–5 token sets are considered. As for the date, texts often contain several mentions, in which case at most 5 of them are extracted.

3.2 Opinion Mining

At the end of 2013 the Russian Federation authorities approved the Law of Education. This event was preceded by a long evaluation of the contents of the Law by the educational community, the Parliament and the Government. It was a tough debate, due to the conflict between the market-driven and the traditional approach to state education financing in the country. We have performed the opinion mining using GMDH [1,3] for the pairs from actor and instance dimensions (Table 2).

It can be seen that the negative attitude was expressed mostly towards a public organization (the Civic Chamber), and not towards the governmental

Table 2. Distribution of opinions on the Law of education

Actor	Instance	Neg	Pos.
The Civic Chamber	Law of Education	72 %	18 %
Ministry of Education	Law of Education	55 %	45 %

Table 3. Contents of the main actor and instance clusters

No.	Actor cluster contents	Instance cluster contents
1	abramov, rukshin, kovaldzhi, kuzminov, higher school of economics, international monetary fund, world bank, russian ministry of education, civic chamber of the russian federation	Strategy, law, standard, lyceums
2	malinetsky, efimov, moscow institute of physics and technology, institute of international programs	Korea, China, Singapore
3	...	...

structures (Ministry of Education). It can be easily explained, because there are many followers of unpopular reforms in the mentioned public organization.

3.3 Clustering

Clustering experiments have been carried out using the *MajorClust* method [7, 14] to group actors and instances in the concept dimension. Earlier, MajorClust proved its efficiency in short texts processing [4,12]. The results are shown in the Table 3.

The first Actor cluster includes person and organization names related to the topics "secondary school", "gifted children", "reforms financing". The second cluster contains person and organization names, related to the topics "high school", "science", "modeling". The first Instance cluster spans instances, related to lawmaking, which are in turn related to the topics "secondary school" and "gifted children". The second cluster includes all the countries, where the authorities have been actively supporting young talents in the recent years (key phrases: "gifted children", "innovations", "Asia").

4 Conclusion and Future Work

The present paper describes a document representation model, based on the tripartite model of ontologies by Peter Mika (Yahoo! Research), which can be useful for the solution of Internet Sociology tasks. The results of a pilot experiment are presented. Information on the deployment of the developed tools is provided. As the future work, we plan to do the following:

- Investigate the behaviour of the proposed model in more detail. To this end:
  - Evaluate the proposed model on a larger corpus;
  - Perform clustering in the actor and instance dimensions;
- Develop the necessary tools that will allow to:
  - Perform document parametrization automatically;
  - Visualize the results so that they can be easily interpreted.

References

1. Alexandrov, M.: Development of general methodology for analysis of public opinion of Internet-community and its application to given topics (authority, economy, corruption, etc.) on the basis of Data/Text Mining tools. Report on State project 84, RPANEPA [rus] (2013)
2. Alexandrov, M., Beresneva, D., Makarov, A.: Dynamic vocabularies as a tool for studying social processes. In: Proceedings of the 6th International Conference on Intelligent Information and Engineering Systems, ITHEA Publishing, vol. 27, pp. 88–92 (2014)
3. Alexandrov, M., Danilova, V., Koshulko, A., Tejada, J.: Models for opinion classification of blogs taken from Peruvian Facebook. In: Proceedings of the 4th International Conference on Inductive Modeling (ICIM-2013), Kyiv, Ukraine Publishing House ITRC-NASU (Ukraine) & Czech Technical University pp. 241–246 (2013)
4. Alexandrov, M., Gelbukh, A., Rosso, P.: An approach to clustering abstracts. In: Montoyo, A., Muñoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 275–285. Springer, Heidelberg (2005)
5. Alexandrov, M., Gelbukh, A., Makagonov, P.: Evaluation of thematic structure of multidisciplinary documents and document flows. In: Proceedings of the 11th International DEXA Workshop (Database and Expert System Applications), pp. 125–129 (2000)
6. Baeza-Yates, R., Ribero-Neto, B.: Modern Information Retrieval. Addison Wesley, Boston (1999)
7. Bishop, C.: Pattern Recognition and Machine Learning. Springer, Berkeley (2006)
8. Danilova, V., Alexandrov, M., Blanco, X.: A survey of multilingual event extraction from text. In: Métais, E., Roche, M., Teisseire, M. (eds.) NLDB 2014. LNCS, vol. 8455, pp. 85–88. Springer, Heidelberg (2014)
9. Danilova V., Popova S.: Socio-political event extraction using a rule-based approach. In: Meersman, R., Panetto, H., Mishra, A., Valencia-Garcia, R., Soares, A.L., Ciuciu, I., Ferri, F., Weichhart, G., Moser, T., Bezzi, M., Chan, H. (eds.) Proceedings of the 13th International Conference on Ontologies, DataBases and Applications of Semantics (ODBASE'2014), vol 8842, pp. 537–546, Springer (2014)
10. Gelbukh, A., Sidorov, G., Guzman-Arenas, A.: Use of a weighted topic hierarchy for document classification. In: Matoušek, V., et al. (eds.) TSD 1999. LNCS (LNAI), vol. 1692, pp. 133–138. Springer, Heidelberg (1999)
11. Manning, C., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
12. Eissen, S.M., Stein, B.: Analysis of clustering algorithms for web-based search. In: Karagiannis, D., Reimer, U. (eds.) PAKM 2002. LNCS (LNAI), vol. 2569, pp. 168–178. Springer, Heidelberg (2002)
13. Mika, P.: Ontologies are us: a unified model of social networks and semantics. J. Web Semant. Sci. Serv. Agents World Wide Web **5**(1), 5–15 (2007)
14. Stein, B., Niggemann, O.: On the nature of structure and its identification. In: Widmayer, P., Neyer, G., Eidenbenz, S. (eds.) WG 1999. LNCS, vol. 1665, p. 122. Springer, Heidelberg (1999)