Discovering and Describing Coherent and Meaningful Topics from a Text Collection

Henry Anaya-Sánchez

IR&NLP-UNED

May 7th, 2013

Content

Motivation

i. The need of information systems and users for analyzing, structuring, and summarizing large collections of text documents according to the **main subject themes that run over the collection documents** (i.e., their **topics**).

ii. Traditional approaches to discover and describe topics based on clustering and Probabilistic Topic Modeling (PTM) are insufficient to always provide ostensible end-users with coherent and meaningful topics.

Motivation

Clustering & PTM are unsupervised learning techniques that has been widely used in the process of topic discovery from documents.

i. Clustering methods aim at generating document groups or clusters, each one representing a different topic.
ii. PTM approaches focus on learning a set of word distributions aimed at generating each document in a collection to represent the topics.

Motivation

However, the obtained clusters/distributions do not necessarily correspond to actual topics of interest:

i. They do not always correlate with human judgements so as to always provide ostensible end-users with **semantically coherent** (interpretable, subject-based) and meaningful (main theme vs. background) topics that summarize the content comprised in a text collection.

Motivation

However, the obtained clusters/distributions do not necessarily correspond to actual topics of interest:

i. They are actually clusters/probability distributions of words with a statistically meaning that sometimes are difficult to interpret and explain by humans since the information they convey in many cases is not at all related to a subject.

Motivation

On the other hand,

i. Clustering methods do not provide descriptions that summarize the clusters' contents (so that users can judge their relevance).
ii. The descriptions provided by PTM approaches are currently limited to list the most probable (frequent) terms under each distribution.

Motivation

This talk presents an approache for discovering topics focused on:

1. how to discover the semantically coherent and meaningful (interpretable, subject-heading like) topics comprised in a text collection, and
2. how to simultaneously provide an appropriate description for each topic so that humans can easily judge its relevance.

A topic discovery methodology based on term pairs

The methodology is closely related to the series of works: FIHC (Fung et al., 2003), CFWS (Li et al., 2008) and the method proposed by Malik and Kender (2006); that aim at obtaining simultaneously both the coverage of a topic and its description by means of a new clustering criterion based on the concept of frequent term set (i.e. a set of terms that co-occur in at least a minimum number of documents in the text collection).

In these works, document clusters and their descriptions are determined by the frequent term sets of the document collection.

A topic discovery methodology based on term pairs

Similarly, our approach relies on highly probable term pairs generated from the collection. However, we use these pairs only as a guide to explore the possible topics of the collection.

Topics and their descriptions are generated from term pairs deemed to be representative of a collection topic.

We introduce the concept of homogeneity of a document set, which is aimed at checking if a set of documents is cohesive enough to represent a topic.

A topic discovery methodology based on term pairs

Input: A set of documents $\mathscr{C} = \{d_1, \ldots, d_n\}$.
Output: The set of topics generated from $\mathscr{C}$ together with their descriptions, $\mathscr{G} = \{(\delta_1, G_1), \ldots, (\delta_m, G_m)\}$.

1. Build the set of term pairs $\mathscr{P}$.
2. Let $\mathscr{G} = \emptyset$.
3. $\pi = \arg\max_{\{t_i,t_j\}\in\mathscr{P}} \mathrm{P}(\{t_i, t_j\}|\mathscr{C})$
4. If $\mathscr{C}|_\pi$ is homogeneous in content then
   (a) $G = Rel(\pi)$
   (b) $\delta = \delta(\pi)$
   (c) $\mathscr{G} = \mathscr{G} \cup \{(\delta, G)\}$
   (d) $\mathscr{C} = \mathscr{C} \setminus G$
5. $\mathscr{P} = \mathscr{P} \setminus \{\pi\}$
6. If $\mathscr{C}\neq\emptyset \wedge \mathscr{P}\neq\emptyset$ then go to Step 3.
7. If $\mathscr{C}\neq\emptyset$ then
   (a) $\mathscr{G} = \mathscr{G} \cup \{(\delta, \{d\}) | d \in \mathscr{C} \wedge \delta$ is the most probable term pair in $d\}$
8. Return $\mathscr{G}$.

A topic discovery methodology based on term pairs

We define the probability of generating a pair of terms $\{t_i, t_j\} \in \mathcal{P}$ from $\mathcal{C}$ as:

$$\mathsf{P}(\{t_i, t_j\}|\mathcal{C}) = \sum_{d \in \mathcal{C}} \mathsf{P}(t_i|d)\mathsf{P}(t_j|d)\mathsf{P}(d|\mathcal{C}) \quad (1)$$

A topic discovery methodology based on term pairs

We propose a novel way to estimate the homogeneity of a set of documents (specifically, for the support set of a given term pair) by analyzing its possible content coverage.

It relies on the concept of pure entropy of a partition $\Theta = \{\Theta_1, \ldots, \Theta_q\}$:

$$H(\Theta) = -\sum_{i=1}^{q} P(\Theta_i|\Theta) \log_2 P(\Theta_i|\Theta) \tag{2}$$

where $P(\Theta_i|\Theta)$ can be estimated as $|\Theta_i| / \sum_{j=1}^{q} |\Theta_j|$.

A topic discovery methodology based on term pairs

Additionally, we provide larger descriptions based on the likelihood ratio score (Dunning, 1993).

This score has been widely used for estimating the correlation of terms with respect to a target topic (Lin and Hovy, 2000; Harabagiu and Lacatusu, 2005).

A topic discovery methodology based on term pairs

For evaluating this approach we have used three benchmark collections: AFP,3 TDT2 version 4.0 and Reuters-21578.4 The Reuters dataset is composed by stories that have been tagged with the attribute TOPICS = YES and include a BODY part.

These collections are different in terms of number of topics, topic sizes, number of dimensions and document distribution.

A topic discovery methodology based on term pairs

Data	Algorithm	Macro-F1	Micro-F1	Overlapping	Itemsets
AFP	FIHC	0.537	0.642	1.0	48084
	FIHC_best_level	0.515	0.603	1.0	48084
	CFWS	0.401	0.463	32.5	53417
	Malik	0.609	0.661	14.6	3047
	Malik_best_level	0.523	0.597	1.1	3047
	Our approach	**0.719**	**0.766**	1.0	134
TDT2	FIHC	0.404	0.515	1.5	40630
	FIHC_best_level	0.278	0.381	1.0	40630
	CFWS	0.095	0.135	27.5	508246
	Malik	0.684	0.748	14.7	5811
	Malik_best_level	0.594	0.689	2.3	5811
	Our approach	**0.868**	**0.901**	1.0	979
Reuters-21578	FIHC	0.174	0.279	1.2	2737
	FIHC_best_level	0.093	0.137	1.0	2737
	CFWS	0.056	0.082	7.5	186021
	Malik	**0.423**	0.526	22.9	2520
	Malik_best_level	0.389	0.513	1.7	2520
	Our approach	0.387	**0.535**	1.0	2126

A topic discovery methodology based on term pairs

Manual topics title/size	Method	Best F1 matching clusters F1/ label/ description
Monica Lewinsky Case/ 969	FIHC	0.87/ *house*/ *white, president, lewinsky, clinton, monica*
	Our approach	0.92/ {*lewinsky, monica*}/ *lewinsky, monica, starr, counsel, grand*
Fossett's Balloon Ride/ 15	FIHC	0.18/ *problem*/ *day, make, year, man, high*
	Our approach	1.00/ {*balloon, world*}/ *fossett, steve, balloon, balloonist, louis*
Current Conflict with Iraq/1486	FIHC	0.92/ *council*/ *security, u.n., iraq, weapon, inspector*
	Our approach	0.85/ {*u.n., iraq*}/ *iraq, u.n., inspector, weapon, iraqi*
Cable Car Crash/ 110	FIHC	0.23/ *force*/ *military, official, death, kill, people*
	Our approach	0.97/ {*cable, car*}/ *cable, marine, italian, car, italy*
Tornado in Florida/ 53	FIHC	0.22/ *central*/ *people, continue, hit, home, kill*
	Our approach	0.95/ {*tornado, florida*}/ *tornado, florida, central, twister, storm*
Oprah Lawsuit/ 70	FIHC	0.47/ *show*/ *talk, make, time, bring, u.s.*
	Our approach	1.00/ {*winfrey, show*}/ *winfrey, oprah, beef, cattle, cow*
LaSalle Boat FOUND!/ 1	FIHC	1.00/ *find, year*/ *authority, expect, large, run*
	Our approach	1.00/ {*lasalle, ship*}/ *divers, lasalle, aimable, explorer, artefact*

Conclusions

- A new methodology for discovering and describing the coherent and meaninful topics comprised in a text collection has been introduced.
- The proposed algorithm provides a novel parameter-less method for discovering the topics from the collection, at the same time that it attaches suitable descriptions to the discovered topics.
- The experiments carried out over TDT2 English corpus, AFP Spanish collection and Reuters-21578 validate our proposal and show significant improvements over existing methods in terms of the standard macro- and micro-averaged F1 measures.