Cognitive Modeling for Topic Expansion
(Short Paper)

Sumant Kulkarni, Srinath Srinivasa, and Rajeev Arora

International Institute of Information Technology Bangalore,
26/C, Electronics City, Bangalore India
{sumant,sri}@iiitb.ac.in, rajeev.arora@iiitb.org
http://www.iiitb.ac.in

Abstract. Expanding a topic into its constituent terms used by a population is an important problem in social space analytics. A topic is defined as a concept representing the semantic "aboutness" of a set of other concepts. What a topic is really "about" may differ across different populations, and discovering this provides sound insights about the population itself. In this paper, we propose an approach for topic expansion, inspired by models from cognitive science – specifically, episodic and semantic memory. We propose an *episodic hypothesis* that asserts a relationship between a topic and other concepts that are collectively about the topic. The canonical form of this algorithm, while shown to produce good results, does not run in interactive response time. We then propose several simplifications over the canonical form to provide interactive response times without reducing too much on the quality of the results. The proposed simplifications also help in creating topical clusters on repositories where there is not enough representation of different terms for the topic.

Keywords: Topic Expansion, Topic Modeling, Text Mining, Social Space Analytics.

1 Introduction

Online social spaces are becoming an integral part of our lives. The need to understand different viewpoints expressed in these spaces has generated interest in social space analytics.

In this paper, we look into the problem of *topic expansion*. The task here is to determine the collective world-view of a population on a given topic. For instance, on Wikipedia we found that the term *Europe* was largely associated with terms pertaining to history and socio-politics, while on an internal Indian corporate blog dataset, *Europe* was associated with terms pertaining to photography and tourism. These collective world-views for a given topic can be of interest to several stakeholders in policy making, strategic planning, marketing and beyond.

This paper uses the 3-layer cognitive model [7,8] for topic expansion. The model is based on proposing an *episodic hypothesis* that makes an assertion about term distribution across episodes to different kinds of semantic associations.

R. Meersman et al. (Eds.): OTM 2013, LNCS 8185, pp. 703–710, 2013.

This paper proposes and tests an episodic hypothesis based on the property of "topical coherence" of an episode. This hypothesis is then tested with human subjects using term distributions from Wikipedia and compared with other approaches for topic modeling.

The proposed hypothesis provided satisfactory results but performance was a major concern. We then experimented with different relaxations to the hypothesis to improve performance without significant changes to the quality of results. Both algorithms are discussed in this paper and a prototype is available over the Internet[1]. A more detailed version of this paper can be found as a technical report [1].

2 Related Literature

The problem of topic expansion is similar to Word Sense Disambiguation (WSD) and Topic Modeling (TM). Automatic solutions for WSD can be classified in two broad categories: Supervised and Unsupervised; both using advancements in machine learning. Supervised WSD approaches [6] use training corpus manually annotated with word senses from a dictionary. Unsupervised approaches, based on the idea of distributional hypothesis [9], do not use them. WSD does not take care of ordering of the terms based on their importance generally.

A Topic modeling algorithm like LDA [2], can be seen to be closely related to topic expansion. Graber et al. [3] extended LDA to use WordNet random walk as an additional source of word generation. Budanitsky et. al [4] found that it is difficult to quantify adhoc semantic senses.

Rachakonda et al. [7,8] use models from cognitive science to develop a 3-layer model for inferring latent semantics. The work presented in this paper is built on top of the 3-layer cognitive model and applied to the problem of topic expansion. Section 3 describes the 3-layer cognitive model in detail. A more detailed and extensive literature survey can be found in [1].

3 3-Layer Cognitive Model

In this section, we briefly introduce the 3-layer cognitive model first proposed in [8], which forms the underlying framework for the topic expansion algorithm. Based on relevant theories in cognitive science, semantic communication in humans is modeled as comprising of three layers called the *analytic* layer, the *episodic* layer and the *linguistic* layer, respectively. Figure 1 depicts this.

The analytic layer is responsible for maintaining our semantic "worldview" in the form of concepts and relationships across them. An axiomatic unidirectional relationship called "aboutness" is defined between a pair of concepts c_1 and c_2 such that the aboutness of c_2 from c_1 is the measure of how much c_2 is about (is relevant to) c_1. The semantics used in the analytic layer is hosted in our

[1] Working prototype of topic expansion can be found at `http://tinyurl.com/topicexpansion`

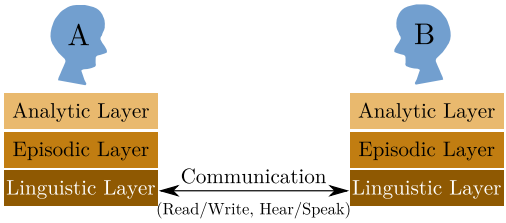

Fig. 1. Semantic communication in humans

semantic memory. Concepts in semantic memory are built from extracting and generalizing on experiences from the *episodic memory*. An episode corresponds to an autobiographical situation from where experiences are recorded.

Our approach towards mining latent semantics focuses on the interplay between semantic and episodic memory. Given a text corpus, the computer acts as the recipient of the semantic communication, and each document in the corpus can be seen as an episode. To generalize on episodic knowledge, we propose an interim data structure called the *co-occurrence graph* that maintains a weighted set of co-occurrences of terms across the corpus. The *co-occurrence graph* represents a primitive and "uncooked" form of analytic layer for machines. Finer semantic constructs are derived by using several algorithms over it. The basis for such constructs are based on proposing relevant *episodic hypotheses* that hypothesize on the patterns of co-occurrences of terms and their relationship with underlying semantics.

An episode is divided into several *occurrence contexts*, and pairs of terms occurring in the same occurrence context are added as edges to the co-occurrence graph. If two terms have co-occurred at least once in the corpus, then there will be an edge between them. We formally define the undirected co-occurrence graph $\mathcal{G}$ as, $\mathcal{G} = (T, C, w)$, where, T is the set of all terms in the given corpus. C is the set of all pair-wise co-occurrence between terms in T. The function $w : C \to \mathbb{N}$ is the corresponding pair-wise co-occurrence count. On top of a given co-occurrence graph, several "primitives" are defined. A pertinent set of primitives used in co-occurrence algorithms of topic expansion are introduced below.

Given a set of terms X, their *focus* $X_\perp$, is the set of terms that co-occur with *all* terms in X. For a term t, its *neighborhood* $N(t)$ is the set of all terms which have co-occurred with it in the corpus along with their co-occurrence weights. We formally define neighborhood as: $N(t) = (T_{N(t)}, C_{N(t)}, w)$, where $T_{N(t)} = \{t\} \cup \{u | u \in T, \{t, u\} \in C\}$, $C_{N(t)} = \{\{t, u\} | u \in T, \{t, u\} \in C\}$ and w is the corresponding edge weight in $\mathcal{G}$ for a given edge.

Given a set of terms X, the neighborhood of its focus $N(X_\perp)$ is defined as: $N(X_\perp) = \bigcap_{x \in X} N(x)$. The set of terms in X are treated as a hypothetical single concept represented by multiple terms when we are talking about neighborhood of $X_\perp$. This allows us to see this neighborhood once again as a star graph. Co-occurrence weights of a given neighbor u from $X_\perp$ is defined as the minimum of all edge weights to u from all the terms in $X_\perp$, corresponding to the multi-set (bag) intersection of co-occurrences. Formally: $w(X_\perp, u) = \min_{x \in X} w(x, u)$.

While the co-occurrence count represents joint occurrences of pairs of terms, from the vantage point of a given term u, we can compute the probability of the occurrence of other terms in the same context as u. This results in a directed graph depicting the "generatability" of a given term u in the context of some other term t. Formally, then the generatability of this co-occurrence $\Gamma_{t\to u}$ is,

$$\Gamma_{t\to u} = \begin{cases} \dfrac{w(t,u)}{\sum_{x\in T_{N(t)}} w(t,x)} & u \neq t, u \in T_{N(t)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

4 Canonical Topic Expansion (CTE)

Topic expansion is a process of unfolding a topic t represented by a given concept, into related concepts, that are collectively *about* t. Terms are linguistic handles for concepts. Each term may represent multiple senses and each sense represents a concept in analytical layer. Hence, sense distinction becomes an implicit part of the topic expansion problem.

We approach topic expansion using the 3-layer cognitive model. First we define an *analytical definition* of topic expansion. Then, we propose an *episodic hypothesis* that asserts observable patterns that indicate topic expansion. Finally, we design a *co-occurrence algorithm* based on the episodic hypothesis that is used for running topic expansion on the given text corpus.

Analytical Definition. For a topic represented by a concept t, a topic expansion $TE(t) = \{c_1, c_2, ..., c_n\}$ is a set of concepts, which collectively display a high *aboutness* for t in our semantic memory. It also makes sense to view topic expansion as an *ordered list* of concepts, based on their individual *aboutness* scores for t. This makes $TE(t)$ as a tuple of terms: $\langle c_1, c_2, ..., c_n\rangle$, where, $A_{c_1}(t) \geq A_{c_2}(t) \geq \ ... \ \geq A_{c_n}(t)$. $A_{c_i}(t)$ is the aboutness score of c_i for t. A concept is always maximally "about" itself, making t the first concept in $TE(t)$.

Episodic Hypothesis *(Topical coherence hypothesis.)* In a long enough conversation (episode) *about* a concept t, terms that are *about* t including t itself tend to be uttered together, compared to terms that are not *about* t.

Co-occurrence Algorithm: The episodic hypothesis is converted into an algorithm over the co-occurrence graph through *cluster generation*, *cluster merging* and *ranking* steps.

Cluster generation: Given a term t with $|N(t)| = k$, we first look for clusters of terms in $N(t)$ containing t. Clustering is based on the generatability of the terms in neighborhood. For k neighbors, this step generates k clusters. The i^{th} run of the clustering algorithm will be based on the topic term t and i^{th} most generatable term. Algorithm 1 explains the i^{th} iteration of the step.

Cluster merging: This step merges redundant clusters based on their similarity calculated as, $O(C_a, C_b) = \frac{|C_a \cap C_b|}{\min(|C_a|,|C_b|)}$ for clusters C_a and C_b. Pairs of clusters having high similarity are merged iteratively till there are no highly similar clusters.

Algorithm 1. Cluster generation with the ith most generatable term

Data: Co-occurrence graph G, topic term t, co-occurring term u_i where u_i is the i^{th} most generatable term in the neighborhood of t, when this algorithm is called for the ith time. u_i is the "sense" of the cluster X.
Result: Cluster C_i of terms containing t, u_i and a subset of terms from $N(t)$
$X \longleftarrow \{t, u_i\}$;
while $N(X_\perp)$ *exists* **do**
  Let $v \in N(X_\perp)$ be the term in $N(X_\perp)$ with the highest generatability $\Gamma_{X_\perp \to v}$;
  $X \longleftarrow X \cup v$;
end
return X;

Ranking: This step orders the terms within each cluster in the order of their importance to the sense of t. This is done using a measure called the *exclusivity* score of a term to t. Exclusivity is the bidirectional generatability score between two terms and is calculated as $E(t_m, t_n) = \Gamma_{t_m \to t_n} \cdot \Gamma_{t_n \to t_m}$. The exclusivity score explains the importance of the term of topic (t) to the terms in the cluster and vice-versa. Detailed explanation on all these steps can be found in [1].

In this algorithm, the *cluster generation step* initially expands the topic for *all* the senses available, even though it might generate redundant clusters. We call this the Canonical Topic Expansion (CTE) algorithm.

4.1 Experimental Evaluation of CTE

Experimental evaluation for the quality of results were performed over a co-occurrence graph constructed over a Wikipedia dump of the year 2006. 25 polysemic terms were chosen as topic terms to evaluate the results. Results of CTE were compared with outcomes of (a) a topic modeling approach based on LDA, and (b) an algorithm for WSD [5]. We generated 3 clusters for each algorithm and asked users to rate the clusters over two factors: *cohesiveness*(C_{val}) and *relatedness*(R_{val}). The C_{val} of a cluster is a measure of how strongly interrelated were the terms in the cluster. The R_{val} is a measure of the relevance of the cluster to the topic term t. It was observed that CTE outperforms the results of both LDA and WSD. The details of the evaluations are found in [1].

The worst case time complexity of CTE is $O(N^5 \log N)$ [1] for corpus with N terms. For *cluster generation* step it is $O(N^3)$. Though it is polynomial, it is still unacceptable for very large N. Co-occurrence graphs tend to be extremely dense (diameter of Wikipedia co-occurrence graph is 4), making average running time close to the worst case.

Another shortcoming of CTE is that, on corpora where document lengths are small, the evidence of terms co-occurring with the *focus* of a large set of candidate terms is unlikely. This form of a co-occurrence requires the existence of a clique in the co-occurrence graph, which is unlikely when document sizes are small. To address these problems – of broadening topic expansion to work on corpora with small document sizes, and make it interactive in response times, we made some simplifications on CTE which are explained in the next section.

5 Interactive Topic Expansion (ITE)

The episodic hypothesis of CTE asserts that the concepts which are "about" a topic tend to cluster together around the topic. This clustering was represented as *cliques* in the co-occurrence algorithm. However, clusters need not always be cliques, and this is more likely in cases where articles in the corpus are not very long. Relaxing the interpretation of a cluster as a clique can broaden the result size of topic expansion as well as reduce the response time significantly.

5.1 Co-occurrence Algorithm for ITE

In CTE, the co-occurrence algorithm assumes that every time a new term is added into a cluster, it is due to its high generatability with *all* other terms already present in the cluster. This assumption is relaxed to give us a set of ITE algorithms. In ITE, initial clusters are formed based on only few *common terms* and then all terms in the focus of the common terms are added to the result set. Different variants of this are proposed, called *Super Fast* (SF) and $Fast_k$ (F_k). ITE algorithms comprise of the same three steps: *Cluster generation, Filtration* and *Ranking*. *Cluster generation* and *Filtration* steps are interleaved.

The *Cluster generation* in F_k methods is performed by first adding $k+1$ more terms to the cluster based on their generatability scores as it was in CTE and then adding all the terms in the focus of the topic term and the other k terms. The term added to the cluster after the topic term is called "key sense". The *k+1* terms are the *anchor terms*. The F_0 is called *SF*.

The generated cluster is given to the *Filtration* step. This step drops the generated cluster if it has overlap higher than the "drop" threshold β with any already generated valid clusters. *Cluster generation* and *Filtration* are performed until all the neighbors are accounted. The clusters surviving filtration are given to *Ranking* step which is same as of CTE. It gives the final ITE clusters.

The *Cluster generation* step in both methods executes faster than in CTE as the worst case complexity of it reduces to $O(N^2)$. The other steps have the same worst case time complexities as of CTE. However, as the worst case is highly improbable in ITE, we claim that ITE performs better than CTE. Following experimental evaluation confirm this. The asymptotic analysis of both CTE and ITE can be found in [1].

5.2 Experimental Results

The experimental evaluation setup for ITE was the same as detailed in section 5.1. The drop threshold (β) was set to 0.5. The performance of ITE was evaluated on four grounds namely: the result similarity with CTE, execution time, C_{val} and R_{val}. The same 25 topic used to evaluate CTE were chosen.

Evaluation of Similarity of Results with CTE: We compared the results generated by both the ITE methods with that of CTE based on the key senses.

For every cluster generated by ITE, we checked whether there is a cluster generated using the same *key sense* term in CTE. If there was such a cluster, we calculated the Jaccard coefficient and overlap of the two clusters. The Jaccard coefficient between two sets is, $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. If no corresponding cluster was found in CTE, the ITE score was 0. We calculated the average of these overlap and Jaccard for 25 terms to get average *Jaccard score* and *Overlap score*. These scores were calculated for 25 iterations starting with *SF* till F_{24}. We also calculated the *speedup* of the ITE compared to CTE for each of the iterations. Speedup signifies the execution time reduction achieved by the ITE compared to CTE. The speedup $S(k)$ for k^{th} iteration with k anchor terms is calculated as $S(k) = \frac{E_c(n) - E_i(n,k)}{E_c(n)}$, where, for n terms, $E_c(n)$ is the execution time of the CTE and $E_i(n,k)$ is the execution time of the ITE in the k^{th} iteration.

We observed that the Jaccard score started with 0.219 and increased with the increase in the number of anchor terms initially. It stabilized at 0.395 after the iteration 19. As the number of anchor terms increased in ITE, the similarity between the ITE and CTE clusters also increased. However, the stabilization of the Jaccard at 0.395 showed that, the ITE can not produce results identical to CTE and has a upper bound in the similarity it can achieve.

The *Overlap Scores* showed a decreasing tendency initially starting from 0.53 however, stabilizing after 19^{th} iteration to 0.47. This shows that initially there were many large ITE clusters for which the corresponding CTE clusters were their *near subsets*. However, as the number of anchor terms increased, the ITE clusters reduced in their size reducing the tendency of CTE clusters being their *near subset*. We also observed that there was a high *speedup* value for all the 25 terms, even though there was a decreasing trend. The *S(25)* still approached 1 (*S(25)* = 0.98) showing the very low execution time of ITE compared to CTE.

User Evaluation. We evaluated with the same 25 terms we used for CTE. Similar to CTE evaluation, we calculated the C_{val} and R_{val} for all three methods: CTE, F_1 and *SF*. We observed that F_1 gave marginally better R_{val} of 3.29, than CTE with a R_{val} of 3.25. *SF* with a R_{val} 3.10 performed worse compared to CTE. We also calculated the average of C_{val} of all terms for CTE, F_1 and *SF* methods. We observed similar trends here as well. F_1 with a C_{val} of 3.20 was the best performer. The CTE with C_{val} 3.19 performed better than *SF* with 3.03.

We observed that the F_1 method performed marginally better than CTE in the user evaluation in both R_{val} and C_{val}. Similarly, *SF* method marginally performed worse compared to CTE. However, there was larger reduction in the execution time (higher *speedup*) as shown earlier.

6 Conclusions and Future Work

The research presented in this paper detailed our explorations into the semantic topic expansion. While a canonical form of the solution was presented, it proved to be too slow to provide an interactive response time. We relaxed the strict interpretations of CTE in the form of ITE, and observed a significant gain in its

response time. It was also shown that, as number of anchor terms increased, ITE results became more similar to CTE results. However, it was also observed that, this similarity gets stabilized at a point which is near 0.5. However, independent user evaluations show that ITE results are considered at least as good, and sometimes even better than the CTE results. The episodic hypothesis based on *topical coherence* stating that *all related terms of a given topic can not co-occur together in one context* seems to hold in human generated corpora.

This work can be extended to look at how *topic maps* can be created from a given text corpus. We can also look into using topic expansion for different forms of query expansion requirements on text corpora.

References

1. Kulkarni, S., Srinivasa, S., Arora, R.: Topic expansion using a term co-occurrence graph. Technical report (2012), `http://rootset.iiitb.ac.in/data/138_topic_expansion_using_tcg.pdf`
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Boyd-Graber, J., Blei, D., Zhu, X.: A topic model for word sense disambiguation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1024–1033 (2007)
4. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist. 32(1), 13–47 (2006)
5. Dorow, B., Widdows, D., Ling, K., Eckmann, J.P., Sergi, D., Moses, E.: Using curvature and markov clustering in graphs for lexical acquisition and word sense discrimination. arXiv:cond-mat/0403693 (2004)
6. Lee, Y.K., Ng, H.T.: An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing - EMNLP 2002, vol. 10, pp. 41–48. ACL, Stroudsburg (2002)
7. Rachakonda, A.R.: A cognitive model for mining latent semantics in unstructured text. In: Proceedings of VLDB PhD Workshop, Istanbul, Turkey (2012)
8. Rachakonda, A.R., Srinivasa, S., Kulkarni, S., Srinivasan, M.S.: Mining analytic semantics from unstructured text. Technical report (2012), `http://rootset.iiitb.ac.in/data/139_mas.pdf`
9. Zellig Harris, S.: Distributional structure. Word 10, 146–162 (1954)