Hybrid Document Indexing with Spectral Embedding

Irina Matveeva
Dept. of Computer Science
University of Chicago
Advisers: Gina-Anne Levow and Partha Niyogi
Motivation

- Content analysis and information analysis requires more linguistic knowledge
- Recent applications such as Summarization, Lexical Entailment, IE work with short paragraphs and sentences and don’t have much disambiguating context
- Bag-of-words does not account for semantic associations between words
- Language Modeling (Ponte and Croft, ‘98) with term translation probabilities (Zhai and Lafferty, ‘01)
- Treat subsets of the vocabulary differently
Overview

- Term and document representation and similarity measure
  - locality
  - collection independence
  - no manually build resources
- Dimensionality reduction
  - spectral embedding
- Multi-level document similarity
  - hybrid representation to support it
Goal: Multi-Level Similarity and Hybrid Document Indexing

- Multi-level document similarity
  - the same people or events
  - semantically related topics
- Vocabulary subsets
  - nouns, Named Entities, verbs
- Define a suitable similarity measure for each subset and compute a representation that supports that measure
  - term matching for Named Entities
  - semantic association between pairs of nouns, similar to language modeling with term translation probabilities
Linguistic Features in Document Indexing

- **Named entities, Noun phrase heads, WordNet synonyms**
- **Synonymy induction using distributional similarity**
  - for nouns (Turney, ‘01)
  - distributional syntactic similarity for verbs (Pantel and Lin, ‘02)
- **Vector space models and spectral methods work well for words** (Schutze ‘97, Widdows ‘03, Matveeva et al. ‘05)
  - for nouns better than for verbs
Overview

• Term and document representation and similarity measure
• Dimensionality reduction
• Multi-level document similarity
Spectral Embedding

• Low-dimensional representation for subsets of the vocabulary
  – Embed words as vectors in a Euclidean space in which cosine is linguistically motivated measure of semantic association

• LSA, PLSA, LDA model term-document association

• Model term-term semantic relations
  – different vocabulary subsets
  – independent of the document collection
Eigenvalue Decomposition

• Obtain a matrix $S$ of pair-wise similarities for all pairs of terms (WordNet, distributional similarity)

• The eigenvalue decomposition of $S$
  
  $$S = U \Sigma U^T$$

  $$S_k = U_k \Sigma_k U_k^T$$ use $k$ largest eigenvalues

  $S_k$ is a product of two matrices

  $$S_k = U_k \Sigma_k U_k^T = U_k \Sigma_k^{1/2} \Sigma_k^{1/2} U_k^T = TT^T$$

• $T$ contains the vector space representation for words

• Cosine between the vectors in $T$ preserves the similarities in $S$
Singular Value Decomposition

- Claim: pair-wise similarities are preserved
- If $S$ is symmetric, its eigenvalue decomposition is the same as SVD
- Singular value decomposition (SVD)
  \[ S = U \Sigma V^T \]
  \[ S_k = U_k \Sigma_k V_k^T \]
- (Eckart and Young) $S_k$ is a matrix $X$ of rank $k$ to minimize
  \[ \|S - X\|_F^2 \]
  where
  \[ \|S - X\|_F^2 = \sum_{ij} (S[i][j] - X[i][j])^2 \]
Spectral Embedding for Nouns with PMI Similarities

- Point-wise mutual information (PMI-IR, Turney ‘00, Terra and Clark ‘03)

\[ \text{PMI}(w_1, w_2) = \log_2 \frac{P(W_1=1, W_2=1)}{P(W_1=1)P(W_2=1)} \]

- Similarities in \( S[i][j] = \text{PMI}(w_i, w_j) \)

- Cosine similarity between the term vectors preserves PMI-based term similarities

- Good performance on the synonymy test (Matveeva et al. ‘05)
Spectral Embedding for Nouns with PMI Similarities - 2

- Vector space embedding for words in which cosine preserves linguistic similarities
  - Good performance on the synonymy test

- Generalizes the ideas of Latent Semantic Analysis, GLSA (Matveeva et al. ‘05)
  - term-term relations
  - locality in computing term associations
  - term associations collection independent
  - different methods of spectral embedding
Synonymy Test

- TOEFL Test 80 questions
- TS 1 Test 50 questions (used 46)
- TS 2 Test 60 questions (used 49)
- Example:
  1. enormously
     a) appropriately, b) uniquely
     c) tremendously, d) decidedly
- Synonymy test was used for LSA
- Better performance than high school students
- PMI-IR (Turney ‘00, Terra and Clark ‘03) outperformed LSA on this test
Synonymy Test: LSA, PMI-IR and Spectral Embedding (GLSA)

- 700,000 documents from the English GigaWord collection to compute PMI
- PMI is computed with a sliding window of a fixed size - notion of locality in the co-occurrence based measures

| Test  | LSA | PMI-IR | GLSA |
|-------|-----|--------|------|
| TOEFL | 0.65| 0.81   | 0.86 |
| TS 1  | ___ | 0.73   | 0.73 |
| TS 2  | ___ | 0.75   | 0.82 |
Overview

- Term and document representation and similarity measure
- Dimensionality reduction
- Multi-level document similarity
Document Representation with Spectral Embedding

• Low-dimensional term vectors $t_j \in \mathbb{R}^k$
  in the space of latent concepts $c_k$
• Documents are linear combinations of term vectors
  $d_i = \sum \alpha(w_j,d_i) \cdot t_j = (c_1, c_2, ..., c_k)$
Hybrid Indexing

- Combine Two Representations
- Tf-idf document vectors
  - \( d_i = (\alpha(w_1,d_i), \alpha(w_2,d_i),..., \alpha(w_n,d_i)) \)
  - \( \alpha(w_j,d_i) = \text{tf}(w_j,d_i) \times \text{idf}(w_j) \)
- Low-dimensional term vectors \( t_j \in \mathbb{R}^k \)
  - in the space of latent concepts \( c_k \)
  - \( d_i = \sum \alpha(w_j,d_i) \times t_j = (c_1, c_2, ..., c_k) \)
- Hybrid Indexing
  - \( d_i = (\alpha(w_1,d_i), \alpha(w_2,d_i),..., \alpha(w_n,d_i), c_1, c_2, ..., c_k) \)
Clustering and classification algorithms are based on the kernel matrix of pair-wise similarities

\[ <d_i,d_j> = \sum_v \alpha(w_v,d_i)*\alpha(w_v,d_j) + \]
\[ (\sum_v \alpha(w_v,d_i) \ t_v)^* (\sum_v \alpha(w_v,d_j) \ t_u) = \]
\[ \sum_v \alpha(w_v,d_i)^* \alpha(w_v,d_i) + \]
\[ \sum_v \sum_u \alpha(w_v,d_i)^* \alpha(w_v,d_j) <t_v,t_u> \]
TDT2 Experiments

• **Collection**
  – Broadcast news from 6 English speaking news sources
  – 10,100 documents assigned to a single topic

• **Question I**
  – Characteristics of the collection
  – Role of Named Entities, similarity measures for verbs

• **Question II**
  – PMI is computed using a large collection with a different word distribution
  – Spectral embedding with PMI has the notion of locality
Document Clustering with Min Squared Residue Algorithm

- Evaluation for cluster \( c_i \) labeled with topic \( t_j \)
  \[
  F(c_i, t_j) = \frac{p*r}{2*(p+r)}
  \]
  \[
  F(C, T) = \sum N_j/N \max F(c_i, t_j)
  \]
- Minimum Squared Residue Co-clustering (Cho et al. ‘04)
  \[
  \beta: \{1, 2, ..., m\} \rightarrow \{1, 2, ..., k\}
  \]
  \[
  \gamma: \{1, 2, ..., k\} \rightarrow \{1, 2, ..., l\}
  \]
  \[
  A(I, J) \text{ is a co-cluster matrix}
  \]
  minimize the sum of the squared differences between the entries in \( A \) and the cluster mean
## TDT2 Clustering

|                | 5-10      | 50-150    | 500-1000  | 1000-5000 |
|----------------|-----------|-----------|-----------|-----------|
| All words      | 0.60(0.09)| 0.80(0.04)| 0.95(0.03)| 0.88(0.07)|
| LSA            | 0.73(0.05)| 0.78(0.05)| 0.98(0.00)| 0.88(0.03)|
| GLSA_local     | 0.81(0.04)| 0.84(0.04)| 0.99(0.00)| 0.90(0.09)|
| GLSA          | 0.64(0.05)| 0.75(0.04)| 0.97(0.00)| 0.93(0.06)|
| only Nouns     | 0.67(0.05)| 0.75(0.03)| 0.97(0.00)| 0.82(0.04)|
| GLSA_Nouns    | 0.80(0.04)| 0.84(0.04)| 0.99(0.00)| 0.92(0.00)|
| Hybrid        | 0.85(0.04)| 0.90(0.03)| 1.00(0.00)| 0.97(0.05)|
# k-NN for Embedded Nouns

| GigaWord  | witness         | testify | prosecutor | trial | testimony | juror | eyewitness |
|-----------|-----------------|---------|------------|-------|-----------|-------|------------|
| TDT2      | witness         | substitution | intimidation | eric  | swoop     | testimony | material   |
| GigaWord  | finance         | fund    | bank       | investment | economy   | crisis | category   |
| TDT2      | finance         | fund    | bank       | investment | economy   | crisis | category   |
| GigaWord  | broadcast       | television | tv         | satellite | abc      | cb     | radio      |
| TDT2      | broadcast       | television | live       | cb     | station   | interview | network    |
| GigaWord  | hearing         | hearing | judge      | voice   | chatter   | sound   | appeal     |
| TDT2      | Hearing         | federal | voice      | sound   | court     | loudness | chant      |
| GigaWord  | Surprise        | announc. | disappoint. | stunning | shock     | reaction | astonishment |
| TDT2      | Surprise        | catch   | demon      | illumination | speciality | bag     | wine       |
| GigaWord  | Rest            | stay    | remain     | keep    | leave     | portion | economy    |
| TDT2      | Rest            | world   | half       | custom  | sound     | lay    | rest       |
Conclusions

- Spectral Embedding preserves linguistically motivated similarities and performs particularly well for nouns
- Named Entities and other characteristic features play an important role
- Multi-level measures of similarity improve document classification and clustering
- Hybrid indexing with spectral embedding provides a principled way to compute a representation that supports that similarity measure
THANK YOU!

Questions?
Comments?