From Frequency to Meaning: Vector Space Models of Semantics

Yifu Huang

School of Computer Science, Fudan University
huangyifu@fudan.edu.cn

COMP620028 Information Retrieval Report, 2013

Outline

Outline

1. Matrices
   - The Word – Context Matrix
   - The Pair – Pattern Matrix
2. Weighting the Elements
   - Positive PMI
3. Smoothing the Matrices
   - Truncated SVD
4. Efficient Comparisons
   - LSH
5. Applications
   - The Word – Context Matrix
   - The Pair – Pattern Matrix

The Word – Context Matrix

- Distributional hypothesis [2]
  - Words that occur in similar contexts tend to have similar meanings
  - "vague", "obscure"
- The context is given by words, phrases, sentences, paragraphs, chapters, documents, or more exotic possibilities, such as sequences of characters or patterns [3]

The Word – Context Matrix

- Distributional hypothesis [2]
  - Words that occur in similar contexts tend to have similar meanings
  - "vague", "obscure"
- The context is given by words, phrases, sentences, paragraphs, chapters, documents, or more exotic possibilities, such as sequences of characters or patterns [3]

The Word – Context Matrix (cont.)

- Example.

Whereof one cannot speak thereof one must be slient

Word	Co-occurrents							
	whereof	one	cannot	speak	thereof	must	be	silent
whereof	0	1	0	0	0	0	0	0
one	1	0	1	0	1	1	0	0
cannot	0	1	0	1	0	0	0	0
speak	0	0	1	0	1	0	0	0
thereof	0	1	0	1	0	1	0	0
must	0	1	0	0	1	0	1	0
be	0	0	0	0	0	1	0	1
silent	0	0	0	0	0	0	1	0

Outline

1. Matrices
   - The Word – Context Matrix
   - The Pair – Pattern Matrix
2. Weighting the Elements
   - Positive PMI
3. Smoothing the Matrices
   - Truncated SVD
4. Efficient Comparisons
   - LSH
5. Applications
   - The Word – Context Matrix
   - The Pair – Pattern Matrix

The Pair – Pattern Matrix

- Extended distributional hypothesis [4]
  - Patterns that co-occur with similar pairs tend to have similar meanings
  - "X solves Y", "Y is solved by X"
- Latent relation hypothesis [5]
  - Pairs of words that co-occur in similar patterns tend to have similar semantic relations
  - "committee:problem", "congress:crisis"

The Pair – Pattern Matrix

- Extended distributional hypothesis [4]
  - Patterns that co-occur with similar pairs tend to have similar meanings
  - "X solves Y", "Y is solved by X"
- Latent relation hypothesis [5]
  - Pairs of words that co-occur in similar patterns tend to have similar semantic relations
  - "committee:problem", "congress:crisis"

The Pair – Pattern Matrix (cont.)

- Example.

Committee finds a solution to problem
Strike is solved by committee
Congress finds a solution to crisis
Congress solves crisis
Civil war is solved by committee

Pair	Pattern		
	X finds a solution to Y	Y is solved by X	X solves Y
committee:problem	1	0	0
strike:committee	0	1	0
congress:crisis	1	0	1
civil war:committee	0	1	0

Outline

Positive PMI

- An alternative to tf-idf which works well for both word-context matrices [6] and pair-pattern matrices [5]
- F be a word-context frequency matrix with n_r rows and n_c columns
- f_{ij} is the number of times that word w_i occurs in the context c_j
- X be the matrix that results when Positive PMI is applied to F
- x_{ij} in X is dened as follows

Positive PMI

- An alternative to tf-idf which works well for both word-context matrices [6] and pair-pattern matrices [5]
- F be a word-context frequency matrix with n_r rows and n_c columns
- f_{ij} is the number of times that word w_i occurs in the context c_j
- X be the matrix that results when Positive PMI is applied to F
- x_{ij} in X is dened as follows

Positive PMI

- An alternative to tf-idf which works well for both word-context matrices [6] and pair-pattern matrices [5]
- F be a word-context frequency matrix with n_r rows and n_c columns
- f_{ij} is the number of times that word w_i occurs in the context c_j
- X be the matrix that results when Positive PMI is applied to F
- x_{ij} in X is dened as follows

Positive PMI

- An alternative to tf-idf which works well for both word-context matrices [6] and pair-pattern matrices [5]
- F be a word-context frequency matrix with n_r rows and n_c columns
- f_{ij} is the number of times that word w_i occurs in the context c_j
- X be the matrix that results when Positive PMI is applied to F
- x_{ij} in X is dened as follows

Positive PMI

- An alternative to tf-idf which works well for both word-context matrices [6] and pair-pattern matrices [5]
- F be a word-context frequency matrix with n_r rows and n_c columns
- f_{ij} is the number of times that word w_i occurs in the context c_j
- X be the matrix that results when Positive PMI is applied to F
- x_{ij} in X is dened as follows

Positive PMI (cont.)

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{n_r} \sum_{j=1}^{n_c} f_{ij}}$$

$$p_{i*} = \frac{\sum_{j=1}^{n_c} f_{ij}}{\sum_{i=1}^{n_r} \sum_{j=1}^{n_c} f_{ij}}$$

$$p_{*j} = \frac{\sum_{i=1}^{n_r} f_{ij}}{\sum_{i=1}^{n_r} \sum_{j=1}^{n_c} f_{ij}}$$

$$\mathrm{pmi}_{ij} = \log\left(\frac{p_{ij}}{p_{i*}p_{*j}}\right)$$

$$x_{ij} = \begin{cases} \mathrm{pmi}_{ij} & \text{if } \mathrm{pmi}_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Outline

Truncated SVD

- An elegant way to improve similarity measurements can be applied to both documents [7] and words [8]
- SVD decomposes X into the product of three matrices $U\Sigma V^T$
- U and V are in column orthonormal form, and Σ is a diagonal matrix of singular values
- If X is of rank r, then Σ is also of rank r, let $k < r$
- Σ_k, the diagonal matrix formed from the top k singular values
- U_k and V_k, the matrices produced by selecting the corresponding columns from U and V
- $\widetilde{X} = U_k \Sigma_k V_k^T$ is the matrix of rank k that best approximates the original matrix X

Truncated SVD

- An elegant way to improve similarity measurements can be applied to both documents [7] and words [8]
- SVD decomposes X into the product of three matrices $U\Sigma V^T$
- U and V are in column orthonormal form, and Σ is a diagonal matrix of singular values
- If X is of rank r, then Σ is also of rank r, let k $<$ r
- Σ_k, the diagonal matrix formed from the top k singular values
- U_k and V_k, the matrices produced by selecting the corresponding columns from U and V
- $\widetilde{X} = U_k \Sigma_k V_k^T$ is the matrix of rank k that best approximates the original matrix X

Truncated SVD

- An elegant way to improve similarity measurements can be applied to both documents [7] and words [8]
- SVD decomposes X into the product of three matrices $U\Sigma V^T$
- U and V are in column orthonormal form, and Σ is a diagonal matrix of singular values
- If X is of rank r, then Σ is also of rank r, let $k < r$
- Σ_k, the diagonal matrix formed from the top k singular values
- U_k and V_k, the matrices produced by selecting the corresponding columns from U and V
- $\widetilde{X} = U_k \Sigma_k V_k^T$ is the matrix of rank k that best approximates the original matrix X

Truncated SVD

- An elegant way to improve similarity measurements can be applied to both documents [7] and words [8]
- SVD decomposes X into the product of three matrices $U\Sigma V^T$
- U and V are in column orthonormal form, and Σ is a diagonal matrix of singular values
- If X is of rank r, then Σ is also of rank r, let $k < r$
- Σ_k, the diagonal matrix formed from the top k singular values
- U_k and V_k, the matrices produced by selecting the corresponding columns from U and V
- $\widetilde{X} = U_k \Sigma_k V_k^T$ is the matrix of rank k that best approximates the original matrix X

Truncated SVD

- An elegant way to improve similarity measurements can be applied to both documents [7] and words [8]
- SVD decomposes X into the product of three matrices $U\Sigma V^T$
- U and V are in column orthonormal form, and Σ is a diagonal matrix of singular values
- If X is of rank r, then Σ is also of rank r, let k < r
- Σ_k, the diagonal matrix formed from the top k singular values
- U_k and V_k, the matrices produced by selecting the corresponding columns from U and V
- $\widetilde{X} = U_k \Sigma_k V_k^T$ is the matrix of rank k that best approximates the original matrix X

Truncated SVD

- An elegant way to improve similarity measurements can be applied to both documents [7] and words [8]
- SVD decomposes X into the product of three matrices $U\Sigma V^T$
- U and V are in column orthonormal form, and Σ is a diagonal matrix of singular values
- If X is of rank r, then Σ is also of rank r, let k < r
- Σ_k, the diagonal matrix formed from the top k singular values
- U_k and V_k, the matrices produced by selecting the corresponding columns from U and V
- $\widetilde{X} = U_k \Sigma_k V_k^T$ is the matrix of rank k that best approximates the original matrix X

Truncated SVD

- An elegant way to improve similarity measurements can be applied to both documents [7] and words [8]
- SVD decomposes X into the product of three matrices $U\Sigma V^T$
- U and V are in column orthonormal form, and Σ is a diagonal matrix of singular values
- If X is of rank r, then Σ is also of rank r, let $k < r$
- Σ_k, the diagonal matrix formed from the top k singular values
- U_k and V_k, the matrices produced by selecting the corresponding columns from U and V
- $\widetilde{X} = U_k \Sigma_k V_k^T$ is the matrix of rank k that best approximates the original matrix X

Outline

1. Matrices
   - The Word – Context Matrix
   - The Pair – Pattern Matrix
2. Weighting the Elements
   - Positive PMI
3. Smoothing the Matrices
   - Truncated SVD
4. Efficient Comparisons
   - LSH
5. Applications
   - The Word – Context Matrix
   - The Pair – Pattern Matrix

LSH

- One general approach to LSH [9] is to “hash” items several times, in such a way that similar items are more likely to be hashed to the same bucket than dissimilar items are
- Denitions of LSH functions include the Min-wise independent function, such as Min-Hashing, that map vectors into short signatures or fingerprints
- After LSH, remained candidate pairs those pairs of signatures that we need to test for similarity

LSH

- One general approach to LSH [9] is to “hash” items several times, in such a way that similar items are more likely to be hashed to the same bucket than dissimilar items are
- Denitions of LSH functions include the Min-wise independent function, such as Min-Hashing, that map vectors into short signatures or fingerprints
- After LSH, remained candidate pairs those pairs of signatures that we need to test for similarity

LSH

- One general approach to LSH [9] is to “hash” items several times, in such a way that similar items are more likely to be hashed to the same bucket than dissimilar items are
- Denitions of LSH functions include the Min-wise independent function, such as Min-Hashing, that map vectors into short signatures or fingerprints
- After LSH, remained candidate pairs those pairs of signatures that we need to test for similarity

Outline

1. Matrices
   - The Word – Context Matrix
   - The Pair – Pattern Matrix
2. Weighting the Elements
   - Positive PMI
3. Smoothing the Matrices
   - Truncated SVD
4. Efficient Comparisons
   - LSH
5. Applications
   - The Word – Context Matrix
   - The Pair – Pattern Matrix

The Word – Context Matrix

- Open Source VSM System: Semantic Vectors [10]
  - Implementing the random projection approach to measuring word similarity
- Word similarity
  - Landauer and Dumais evaluated this approach with 80 multiple-choice synonym questions from the Test of English as a Foreign Language (TOEFL), achieving human-level performance [8]
- Word clustering
  - These algorithms are able to discover different senses of polysemous words, generating different clusters for each sense [6]

The Word – Context Matrix

- Open Source VSM System: Semantic Vectors [10]
  - Implementing the random projection approach to measuring word similarity
- Word similarity
  - Landauer and Dumais evaluated this approach with 80 multiple-choice synonym questions from the Test of English as a Foreign Language (TOEFL), achieving human-level performance [8]
- Word clustering
  - These algorithms are able to discover different senses of polysemous words, generating different clusters for each sense [6]

The Word – Context Matrix

- Open Source VSM System: Semantic Vectors [10]
  - Implementing the random projection approach to measuring word similarity
- Word similarity
  - Landauer and Dumais evaluated this approach with 80 multiple-choice synonym questions from the Test of English as a Foreign Language (TOEFL), achieving human-level performance [8]
- Word clustering
  - These algorithms are able to discover different senses of polysemous words, generating different clusters for each sense [6]

The Word – Context Matrix (cont.)

- Automatic thesaurus generation
  - Creating and maintaining such lexical resources is labour intensive, so it is natural to wonder whether the process can be automated to some degree [11]
- Context-sensitive spelling correction
  - These confusions cannot be detected by a simple dictionary-based spelling checker; they require context-sensitive spelling correction [12]
- Semantic role labeling
  - Word-context matrices can reliably predict the semantic frame to which an unknown lexical unit refers, with good levels of accuracy [13]

The Word – Context Matrix (cont.)

- Automatic thesaurus generation
  - Creating and maintaining such lexical resources is labour intensive, so it is natural to wonder whether the process can be automated to some degree [11]
- Context-sensitive spelling correction
  - These confusions cannot be detected by a simple dictionary-based spelling checker; they require context-sensitive spelling correction [12]
- Semantic role labeling
  - Word-context matrices can reliably predict the semantic frame to which an unknown lexical unit refers, with good levels of accuracy [13]

The Word – Context Matrix (cont.)

- Automatic thesaurus generation
  - Creating and maintaining such lexical resources is labour intensive, so it is natural to wonder whether the process can be automated to some degree [11]
- Context-sensitive spelling correction
  - These confusions cannot be detected by a simple dictionary-based spelling checker; they require context-sensitive spelling correction [12]
- Semantic role labeling
  - Word-context matrices can reliably predict the semantic frame to which an unknown lexical unit refers, with good levels of accuracy [13]

Outline

The Pair – Pattern Matrix

- Open Source VSM System: Latent Relational Analysis in S-Space [14]
  - Pattern frequencies are counted and then smoothed using SVD
- Relational similarity
  - Turney evaluated this approach to relational similarity with 374 multiple-choice analogy questions from the SAT college entrance test, achieving human-level performance [15]
- Relational clustering
  - The representative pairs to automatically generate multiple-choice analogy questions, in the style of SAT analogy questions [16]

The Pair – Pattern Matrix

- Open Source VSM System: Latent Relational Analysis in S-Space [14]
  - Pattern frequencies are counted and then smoothed using SVD
- Relational similarity
  - Turney evaluated this approach to relational similarity with 374 multiple-choice analogy questions from the SAT college entrance test, achieving human-level performance [15]
- Relational clustering
  - The representative pairs to automatically generate multiple-choice analogy questions, in the style of SAT analogy questions [16]

The Pair – Pattern Matrix

- Open Source VSM System: Latent Relational Analysis in S-Space [14]
  - Pattern frequencies are counted and then smoothed using SVD
- Relational similarity
  - Turney evaluated this approach to relational similarity with 374 multiple-choice analogy questions from the SAT college entrance test, achieving human-level performance [15]
- Relational clustering
  - The representative pairs to automatically generate multiple-choice analogy questions, in the style of SAT analogy questions [16]

The Pair – Pattern Matrix (cont.)

- Relational classification
  - Taint:poison is classified as strength (poisoning is stronger than tainting) and assess:review is classified as enablement (assessing is enabled by reviewing) [17]
- Relational search
  - A query for a relational search engine is "list all X such that X causes cancer". In this example, the relation, cause, and one of the terms in the relation, cancer, are given by the user, and the task of the search engine is to find terms that satisfy the user's query [18]
- Analogical mapping
  - With a pair-pattern matrix, we can solve proportional analogies by selecting the choice that maximizes relational similarity [5]

The Pair – Pattern Matrix (cont.)

- Relational classification
  - Taint:poison is classified as strength (poisoning is stronger than tainting) and assess:review is classified as enablement (assessing is enabled by reviewing) [17]
- Relational search
  - A query for a relational search engine is “list all X such that X causes cancer". In this example, the relation, cause, and one of the terms in the relation, cancer, are given by the user, and the task of the search engine is to find terms that satisfy the user's query [18]
- Analogical mapping
  - With a pair-pattern matrix, we can solve proportional analogies by selecting the choice that maximizes relational similarity [5]

The Pair – Pattern Matrix (cont.)

- Relational classification
  - Taint:poison is classified as strength (poisoning is stronger than tainting) and assess:review is classified as enablement (assessing is enabled by reviewing) [17]
- Relational search
  - A query for a relational search engine is “list all X such that X causes cancer". In this example, the relation, cause, and one of the terms in the relation, cancer, are given by the user, and the task of the search engine is to find terms that satisfy the user's query [18]
- Analogical mapping
  - With a pair-pattern matrix, we can solve proportional analogies by selecting the choice that maximizes relational similarity [5]

References I

- [1] From Frequency to Meaning: Vector Space Models of Semantics. JAIR. 2010.
- [2] The Distributional Hypothesis. IJL. 2008.
- [3] The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. PhD Thesis. 2006.
- [4] DIRT – Discovery of Inference Rules from Text. SIGKDD. 2001.
- [5] The latent relation mapping engine: Algorithm and experiments. JAIR. 2008.
- [6] Discovering word senses from text. SIGKDD. 2002.
- [7] Indexing by latent semantic analysis. JASIS. 1990.

References II

- [8] A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. PR. 1997.
- [9] On the resemblance and containment of documents. SEQUENCES. 1997.
- [10] Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. LREC. 2008.
- [11] Improvements in automatic thesaurus extraction. SIGLEX. 2002.
- [12] Contextual spelling correction using latent semantic analysis. ANLP. 1997.
- [13] Automatic induction of FrameNet lexical units. EMNLP. 2008.

References III

- [14] Measuring Semantic Similarity by Latent Relational Analysis. CL. 2006.
- [15] Similarity of semantic relations. CL. 2006.
- [16] Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions. ACL-HLT. 2008.
- [17] VerbOcean: Mining the web for ne-grained semantic verb relations. EMNLP. 2004.
- [18] Relational web search. TR. 2006.