Information Retrieval
Vector Model

Referent: Anand Mishra

Seminar für Computerlinguistik
Ruprecht-Karls-Universität Heidelberg
Hauptseminar “Information Retrieval”
Leitung: PD Dr. Karin Haenelt

05.02.2007
Plan

- Revision
- Implementation (Python)
- Demonstration
- Evaluation
Definition

An information retrieval model is a quadruple \((D, Q, F, R(q_i, d_j))\) where

1. \(D\) is a set of logical views/representations for the documents in the collection
2. \(Q\) is a set of logical views/representations for queries
3. \(F\) is a framework for modeling document representations, queries and their relationships
4. \(R(q_i, d_j)\) is a ranking function defining ordering among the documents \(d_j\) with regard to the query \(q_i\)

(D,Q,F,R(q_i, d_j))

Document-representation-1

- Set of Documents: TIME Corpus (English)
- 423 Documents, ca. 1.5 MB
- `documentReader()`: Corpus ⇒ {docId:content}
- `tokenizer()`: content ⇒ iterator over tokens
- 2,49,069 tokens
- `index-terms`: distinct words in a document = 20,856
- `filterStopWords` ⇒ 20,520
- `stemmer.porter` ⇒ 13,725
- System has **13,725 index-terms**
- Set of index-terms: \( T = \{t_i\}_{i=1}^{n} \) (n=13,725)
Given the document $d_j$ and a term $t_i$ belonging to it, how well $t_i$ represents $d_j$?

Measuring relevance of an index-term $t_i$ for a Document $d_j$ by associating a number (weight $w_{i,j}$) to the term $t_i$: $w_{i,j} t_i$

- $w_{i,j} \geq 0 \ \forall i, j$
- $w_{i,j} = 0$ if $t_i \notin d_j$

Document $d_j = \{ w_{1,j} t_1, w_{2,j} t_2, w_{3,j} t_3, \ldots w_{n,j} t_n \}$

Given $t_1, t_2$, if $w_{1,j} > w_{2,j}$ then, relevance($t_1$) $>$ relevance($t_2$) for the Document $d_j$
Document-representation in Vector Space Model

- Document $d_j = \{w_{1,j}t_1, w_{2,j}t_2, w_{3,j}t_3, \ldots, w_{n,j}t_n\}$
- Let us consider an $n$-dimentional vector space $V_n$
- A vector $\vec{d} \in V_n$ is represented as:
  $\vec{d} = a_1\vec{u}_1 + a_2\vec{u}_2 + a_3\vec{u}_3 + \ldots + a_n\vec{u}_n$
  where $\vec{u}_i$ is a unit-vector on $i^{th}$-axis i.e. $|\vec{u}_i| = 1$
- A vector $\vec{d}_j \in V_n$ can be represented as:
  $\vec{d}_j = a_{1,j}\vec{u}_1 + a_{2,j}\vec{u}_2 + a_{3,j}\vec{u}_3 + \ldots + a_{n,j}\vec{u}_n$
- Framework $F$
  Take $a_{i,j} = w_{i,j}$ then the vector $\vec{d}_j$ represents the document $d_j$
- Thus $d_j \overset{F}{\Rightarrow} \vec{d}_j = \sum_{i=1}^{n} w_{i,j}\vec{u}_i$
  with $\{\vec{u}_i\}_{i=1}^{n}$ are $n$ orthogonal-unit-vectors
\[ d_j \xrightarrow{F} \vec{d}_j = \sum_{i=1}^{n} w_{i,j} \vec{u}_i \]

Calculating \( w_{i,j} \)

- Similar documents closer together, different ones separated
- **intra-cluster similarity**
  quantified by measuring raw-frequency \((freq_{i,j})\) of a term \( t_i \) in document \( d_j \)
  \[ f_{i,j} = \frac{freq_{i,j}}{\max_i freq_{i,j}} \] Normalized frequency
- **inter-cluster dissimilarity**
  quantified by inverse of the frequency of a term \( t_i \) among the documents in the collection \((idf_i)\)
  \[ idf_i = \log \frac{M}{m_i} \]
  where \( M\): total no. of docs. \( m_i\) no of docs. in which \( t_i \) appears
  \[ w_{i,j} = f_{i,j} \times idf_i \]
(D,Q,F,R(q_i, d_j))

Document-representation in Vector Space Model

- indexer() ⇒ \{term_i:{docId_j:freq_{i,j}}\}
- calculate_tf_idf_w() ⇒ \{term_i:{docId_j:(freq_{ij},tf_{ij},idf_i,w_{ij})}\}
- query()
(D, Q, F, R(q_i, d_j))

Ranking

- Quantifying how similar is a document \( d_j \) to query \( q \)
- Correlation between the vectors \( \vec{d}_j \) and \( \vec{q} \)
- \( \cos(\theta) \) where \( \theta = \) angle between the vectors \( \vec{d}_j \) and \( \vec{q} \)
- \( \text{sim}_0(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \)
  Here \( 0 \leq \cos(\theta) \leq 1 \)
- \( \text{sim}_1(d_j, q) = \vec{d}_j \cdot \vec{q} \) Inner Product
- \( \text{ranking()} \Rightarrow \{ \text{rank:docId}_j \} \)
Evaluation
Recall, Precision

- Collection of documents (TIME.ALL)
- Set of queries (TIME.QUE)
- Set of relevant documents provided by specialists (TIME.REL)
- Similarity between set of documents retrieved and TIME.REL

**Recall** is the fraction of the relevant documents (the set $R$) which has been retrieved $\frac{|R_a|}{|R|}$

**Precision** is the fraction of the retrieved documents (the set $A$) which is relevant $\frac{|R_a|}{|A|}$

*Baeza-Yates, Ribeiro-Neto: Modern Information Retrieval, 1999*
References


   http://kontext.fraunhofer.de/haenelt/kurs/InfoRet/


5. www.python.org