Information Retrieval

Retrieval Models III: Pagerank, other models (cont.)

Qianni Zhang

Retrieval Models III: Pagerank, others (cont.)

Roadmap of the next two lectures:

- Pagerank
- Set based model
- Fuzzy set model
- Extended boolean model
- Generalised vector space

Fuzzy set model (vagueness)

Premises

- Documents and queries are represented through sets of keywords, therefore the matching between them is vague
  - Keywords cannot completely describe the user's information need and the doc's main theme

- For each query term (keyword)
  - Define a fuzzy set and that each doc has a degree of membership (0~1) in the set

Fuzzy set model (vagueness)

Theory

- Framework for representing classes (sets) whose boundaries are not well defined
  - Key idea is to introduce the notion of a degree of membership associated with the elements of a set
  - This degree of membership varies from 0 to 1 and allows modelling the notion of marginal membership
    - 0 →no membership
    - 1 →full membership
  - Thus, membership is now a gradual instead of abrupt
    - Not as conventional Boolean logic

Here we will define a fuzzy set for each query (or index) term, thus each doc has a degree of membership in this set.

Fuzzy set model (vagueness)

Definition

- A fuzzy subset A (in universe U) whose boundaries are not well defined (e.g., tall, nice, relevant)
  - Is characterized by a membership function $\mu_A : U \mapsto [0, 1]$
  - Which associates with each element u of U number $\mu_A(u)$ in the interval [0, 1]
- Query term = fuzzy set
- Document = has a membership (between 0 and 1) to that set

- Let A and B be two fuzzy subsets of U. Also, let $\overline{A}$ be the complement of A. Then,
  - Complement $\mu_{\overline{A}}(u) = 1 - \mu_A(u)$
  - Union $\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$
  - Intersection $\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$

Fuzzy set model (vagueness)

Fuzzy set model

- Set-membership function: $\mu_A : U \rightarrow [0; 1]$ where A is a term and U is a set of documents (thesaurus).
  - sailing = { (0.9, d1), (0.8, d2) }
  - boats = { (0.5, d1), (0.8, d2) }
- Membership based on tf, idf, or correlation matrix C:
  - $C_{i,j} := \dfrac{n_{i,j}}{n_i + n_j - n_{i,j}}$
  - $n_{i,j}$: Number of documents in which t_i and t_j occur
  - n_i : Number of documents in which t_i occurs
  - $\mu_i(d) := 1 - \prod_{t_j \in d} (1 - C_{i,j})$
- The weight of a document in the set of term t_i is computed as the disjunction of all document terms related to term t_i .

Fuzzy set model (vagueness)

Fuzzy set model

$$\mu_i(d) := 1 - \prod_{t_j \in d} (1 - C_{i,j})$$

- Document d belongs to fuzzy set (term) t_i if its own terms (the t_js) are related to t_i :
  - one term t_j in document d very related to t_i ($C_{ij} \sim 1$):
    - membership of d to fuzzy set t_i close to 1 ($\mu_i(d) \sim 1$)
  - no term t_j in document d related to t_i (all $C_{ij} \sim 0$):
    - membership of d to fuzzy set t_i close to 0 ($\mu_i(d) \sim 0$)

Fuzzy set model (vagueness)

Summary

- Advantages
  - The correlations among index terms are considered
  - Degree of relevance between queries and docs can be achieved

- Disadvantages
  - Fuzzy IR models have been discussed mainly in the literature associated with fuzzy theory
  - Do not consider the frequency (or counts) of a term in a document or a query

Extended Boolean model

Motivation

- Extend the Boolean model with the functionality of partial matching and term weighting
  - E.g.: in Boolean model, for the query $q = t_x \wedge t_y$, a doc contains either t_x or t_y is as irrelevant as another doc which contains neither of them
  - How about the disjunctive query $q = t_x \vee t_y$
- Combine Boolean query formulations with characteristics of the vector model
  - Term weighting
  - Algebraic distances for similarity measures

A ranking can be obtained

Extended Boolean model

Introduction

- Deal with problems:
  - t_1 AND t_2 (too few documents retrieved) –> affect recall
  - t_1 OR t_2 (too many documents retrieved) –> affect precision
- Use weights $d = (w_1,\ w_2)$ for terms t_1 AND t_2
- OR-queries:

$$R(d,q) = \sqrt{\frac{w_1^2 + w_2^2}{2}}$$

- AND-queries:

$$R(d\,|\,q) = 1 - \sqrt{\frac{(1 - w_1)^2 + (1 - w_2)^2}{2}}$$

$\in [0, 1]$

Extended Boolean model

Extend the idea to *m* terms, *p* norm

- OR-queries : $q_{or} = t_1 \vee^p t_2 \vee^p \ldots \vee^p t_m$

$$R(d,q) = \left(\frac{w_1^p + w_2^p + \ldots + w_m^p}{m}\right)^{1/p}$$

- AND-queries: $q_{and} = t_1 \wedge^p t_2 \wedge^p \ldots \wedge^p t_m$

$$R(d \mid q) = 1 - \left(\frac{(1-w_1)^p + (1-w_2)^p + \ldots + (1-w_m)^p}{m}\right)^{1/p}$$

Extended Boolean model

Example 1, $p = 2$

- t_1 and t_2 in document ($w_1 = 1$ and $w_2 = 1$):
  - OR-query:
  - AND-query:
- t_1 and t_2 not in document ($w_1 = 0$ and $w_2 = 0$):
  - OR-query:
  - AND-query:
- t_1 in document and t_2 not in document ($w_1 = 1$ and $w_2 = 0$):
  - OR-query:
  - AND-query:

Extended Boolean model

Example 2

- Consider the query $q=(t_1 \wedge t_2) \vee t_3$.
- The similarity $R(d,q)$ between a document d and this query is then computed as

$$R(d,q) = \left(\frac{\left(1 - \left(\frac{(1-w_1)^p + (1-w_2)^p}{2} \right)^{1/p} \right)^p + w_3{}^p}{2} \right)^{1/p}$$

Extended Boolean model

Advantages

- A hybrid model including properties of both the set theoretic models and the algebraic models
- That is, relax the Boolean algebra by interpreting Boolean operations in terms of algebraic distances
  - By varying the parameter p between 1 and infinity, we can vary the p-norm ranking behaviour from that of a vector-like ranking to that of a fuzzy logic-like ranking
  - Have the possibility of using combinations of different values of the parameter p in the same query request

Disadvantages

- Assumes mutual independence of index terms

Generalized Vector Space Model

Premise

- Classic models enforce independence of index terms
- For the vector space model (VSM)
  - Set of term vectors $\{\vec{t}_1, \vec{t}_2, \ldots., \vec{t}_N\}$ are linearly independent and form a basis for the subspace of interest
  - Frequently, it means pair-wise orthogonality

$$\forall_{i,j} \Rightarrow \vec{t}_i \cdot \vec{t}_j = 0$$

- Wong et al. proposed an alternative interpretation
  - The index terms are linearly independent, but not pair-wise orthogonal
  - Generalized Vector Space Model (GVSM)

Wong, S. K. M.; Wojciech Ziarko, Patrick C. N. Wong (1985-06-05), *Generalized vector spaces model in information retrieval*, SIGIR ACM

Generalized Vector Space Model

Key idea

- Index term vectors form the basis of the space are not orthogonal and are represented in terms of smaller components (minterms)
- Term to term correlations considered, which deprecate the pair-wise orthogonality assumption.

Notations

- $\{t_1, t_2, \ldots, t_N\}$: the set of all terms
- $w_{i,k}$: the weight associated with $[t_i, d_k]$ (d_k is the k_{th} doc.)
- **Minterms**: binary indicators (0 or 1) of all patterns of occurrence of terms within documents
  - Each represents one kind of co-occurrence of index terms in a specific document

Generalized Vector Space Model

- For a document d_k and a query q the similarity function now becomes:

$$sim(d_k, q) = \frac{\sum_{j=1}^{N} \sum_{i=1}^{N} w_{i,k} * w_{j,q} * t_i \cdot t_j}{\sqrt{\sum_{i=1}^{N} w_{i,k}^2} * \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

- where t_i and t_j are now vectors of a 2^N dimensional space.

- Term correlation $t_i \bullet t_j$ can be implemented in several ways.
- As an example Wong et al. use as input to their algorithm the term occurrence frequency matrix obtained from automatic indexing and the output is term correlation between any pair of index terms.

Generalized Vector Space Model

Representation of minterms

A new space,
where each term vector t_i $i=\{1,2,\ldots,N\}$
is expressed as a linear combination
of 2^N vectors m_r where $r = 1 \ldots 2^N$

Pairwise orthogonal vectors $\vec{m}_i$ associated with minterms m_i as the basis for the generalized vector space

Points to the docs where only index terms t_1 and t_2 co-occur and the other index terms disappear

Point to the docs containing all the index terms

$$
\begin{aligned}
m_1 &= (0,0,\ldots.,0) \\
m_2 &= (1,0,\ldots.,0) \\
m_3 &= (0,1,\ldots.,0) \\
m_4 &= (1,1,\ldots.,0) \\
m_5 &= (0,0,1,..,0) \\
&\ldots. \\
m_{2^N} &= (1,1,\ldots.,1)
\end{aligned}
\qquad \Longleftrightarrow \qquad
\begin{aligned}
\vec{m}_1 &= (1,0,0,0,0,\ldots.,0) \\
\vec{m}_2 &= (0,1,0,0,0,\ldots.,0) \\
\vec{m}_3 &= (0,0,1,0,0,\ldots.,0) \\
\vec{m}_4 &= (0,0,0,1,0,\ldots.,0) \\
\vec{m}_5 &= (0,0,0,0,1,\ldots.,0) \\
&\ldots. \\
\vec{m}_{2^N} &= (0,0,0,0,0,\ldots.,1)
\end{aligned}
$$

2^N minterms　　　　2^N minterm vectors

Generalized Vector Space Model

- Minterm vectors are pairwise orthogonal.
- But, this does not mean that the index terms are independent
  - Each minterm specifies a kind of dependence among index terms
  - That is, the co-occurrence of index terms inside docs in the collection induces dependencies among these index terms

Generalized Vector Space Model

- The vector associated with the term t_i is represented by summing up all minterms containing it and normalizing

$$\vec{t}_i = \frac{\sum_{\forall r, g_i(m_r)=1} c_{i,r} \vec{m}_r}{\sqrt{\sum_{\forall r, g_i(m_r)=1} c_{i,r}^2}} = \sum_{\forall r, g_i(m_r)=1} \hat{c}_{i,r} \vec{m}_r$$

$$\text{where } \hat{c}_{i,r} = \frac{c_{i,r}}{\sqrt{\sum_{\forall r, g_i(m_r)=1} c_{i,r}^2}}$$

$$c_{i,r} = \sum_{d_j | g_l(\vec{d}_j) = g_l(m_r), \text{for all } l} w_{i,j}$$

The weight associated with the pair $[t_i, m_r]$

Summing up the weights of the term t_i in all the docs which have a term occurrence pattern given by m_r.

All the docs whose term co-occurrence relation (pattern) can be represented as (exactly coincide with that of) minterm m_r

$g_i(m_r)$ indicates the index term t_i is in the minterm m_r

Generalized Vector Space Model

- Based on the above, we have:

$$\vec{t}_i = \sum\nolimits_{\forall r, g_i(m_r)=1} \hat{c}_{i,r} \vec{m}_r$$

$$\vec{t}_i \cdot \vec{t}_j = \sum_{\forall r | g_i(m_r)=1 \wedge g_j(m_r)=1} \hat{c}_{i,r} \times \hat{c}_{j,r}$$

$$d_k = \sum\nolimits_{\forall i} w_{i,k} \vec{t}_i$$

$$q = \sum\nolimits_{\forall j} w_{j,q} \vec{t}_j$$

Generalized Vector Space Model

Example

- Suppose that the system has 12 documents (d_1-d_{12}) and 4 terms (t_1-t_4)

  d_1 =(2, 1, 0, 0), d_2 =(5, 1, 0, 0), d_3 =(1, 1, 1, 1),
  d_4 =(0, 0, 2, 2), d_5 =(0, 1, 1, 2), d_6 =(0, 0, 1, 1),
  d_7 =(0, 0, 1, 0), d_8 =(1, 1, 0, 0), d_9 =(2, 1, 1, 1),
  d_{10} =(0, 2, 2, 2). d_{11} =(1, 0, 2, 0), d_{12} =(0,0, 2,1).

- 6 minterms are used as independent vectors to form a base

  m_1=(1, 1, 0, 0), m_2=(1, 1, 1, 1), m_3=(0, 0, 1, 1),
  m_4=(0, 1, 1, 1), m_5=(0, 0,1, 0), m_6=(1, 0, 1, 0).

Generalized Vector Space Model

Example

- Independent vectors:

$$\vec{m}_1 = (1,0,0,0,0,0) \quad \vec{m}_2 = (0,1,0,0,0,0)$$
$$\vec{m}_3 = (0,0,1,0,0,0) \quad \vec{m}_4 = (0,0,0,1,0,0)$$
$$\vec{m}_5 = (0,0,0,0,1,0) \quad \vec{m}_6 = (0,0,0,0,0,1)$$

- $\vec{m}_i$ represents minterm m_i.
- Each pair of $\vec{m}_i$ and $\vec{m}_j$ is orthogonal. (dot product=0)

Generalized Vector Space Model

Example

The four keywords t_1, t_2, t_3, and t_4 are represented by a combination of the independent vectors.

$$\vec{t}_1 = (c_{1,1}\vec{m}_1 + c_{1,2}\vec{m}_2 + c_{1,3}\vec{m}_3 + c_{1,4}\vec{m}_4 + c_{1,5}\vec{m}_5 + c_{1,6}\vec{m}_6) / C$$

where

$$c_{1,1} = w_{1,1} + w_{1,2} + w_{1,8} = 2 + 5 + 1 = 8$$

$$c_{1,2} = w_{1,3} + w_{1,9} = 1 + 2 = 3$$

$$c_{1,3} = w_{1,4} + w_{1,6} + w_{1,12} = 0 + 0 + 0 = 0$$

$$c_{1,4} = w_{1,5} + w_{1,10} = 0 + 0 = 0$$

$$c_{1,5} = w_{1,7} = 0, \qquad c_{1,6} = w_{1,11} = 1$$

$$C = \sqrt{(c_{1,1}^2 + c_{1,2}^2 + c_{1,3}^2 + c_{1,4}^2 + c_{1,5}^2 + c_{1,6}^2)}$$

Generalized Vector Space Model

Example

The four keywords t_1, t_2, t_3, and t_4 are represented by a combination of the independent vectors.

$$\vec{t}_2 = (c_{2,1}\vec{m}_1 + c_{2,2}\vec{m}_2 + c_{2,3}\vec{m}_3 + c_{2,4}\vec{m}_4 + c_{2,5}\vec{m}_5 + c_{2,6}\vec{m}_6) / C$$

where

$$c_{2,1} = w_{2,1} + w_{2,2} + w_{2,8} = 1 + 1 + 1 = 3$$

$$c_{2,2} = w_{2,3} + w_{2,9} = 1 + 1 = 2$$

$$c_{2,3} = w_{2,4} + w_{2,6} + w_{2,12} = 0 + 0 + 0 = 0$$

$$c_{2,4} = w_{2,5} + w_{2,10} = 1 + 2 = 3$$

$$c_{2,5} = w_{2,7} = 0, \qquad c_{2,6} = w_{2,11} = 0$$

$$C = \sqrt{(c_{2,1}^2 + c_{2,2}^2 + c_{2,3}^2 + c_{2,4}^2 + c_{2,5}^2 + c_{2,6}^2)}$$

Generalized Vector Space Model

Example

The four keywords t_1, t_2, t_3, and t_4 are represented by a combination of the independent vectors.

$$\vec{t}_3 = (c_{3,1}\vec{m}_1 + c_{3,2}\vec{m}_2 + c_{3,3}\vec{m}_3 + c_{3,4}\vec{m}_4 + c_{3,5}\vec{m}_5 + c_{3,6}\vec{m}_6) / C$$

where

$$c_{3,1} = w_{3,1} + w_{3,2} + w_{3,8} = 0$$

$$c_{3,2} = w_{3,3} + w_{3,9} = 1 + 1 = 2$$

$$c_{3,3} = w_{3,4} + w_{3,6} + w_{3,12} = 2 + 1 + 2 = 5$$

$$c_{3,4} = w_{3,5} + w_{3,10} = 1 + 2 = 3$$

$$c_{3,5} = w_{3,7} = 1, \qquad c_{3,6} = w_{3,11} = 2$$

$$C = \sqrt{(c_{3,1}^2 + c_{3,2}^2 + c_{3,3}^2 + c_{3,4}^2 + c_{3,5}^2 + c_{3,6}^2)}$$

Generalized Vector Space Model

Example

The four keywords t_1, t_2, t_3, and t_4 are represented by a combination of the independent vectors.

$$\vec{t}_4 = (c_{4,1}\vec{m}_1 + c_{4,2}\vec{m}_2 + c_{4,3}\vec{m}_3 + c_{4,4}\vec{m}_4 + c_{4,5}\vec{m}_5 + c_{4,6}\vec{m}_6)/C$$

where

$$c_{4,1} = w_{4,1} + w_{4,2} + w_{4,8} = 0$$

$$c_{4,2} = w_{4,3} + w_{4,9} = 1 + 1 = 2$$

$$c_{4,3} = w_{4,4} + w_{4,6} + w_{4,12} = 2 + 1 + 1 = 4$$

$$c_{4,4} = w_{4,5} + w_{4,10} = 2 + 2 = 4$$

$$c_{4,5} = w_{4,7} = 0, \qquad c_{4,6} = w_{4,11} = 0$$

$$C = \sqrt{(c_{4,1}^2 + c_{4,2}^2 + c_{4,3}^2 + c_{4,4}^2 + c_{4,5}^2 + c_{4,6}^2)}$$

Generalized Vector Space Model

Example

- d_k the documents are converted from a vector of length 4 into a vector of length 6.

$$d_k = \sum_{\forall i} w_{i,k} \vec{t}_i$$

- Provide the new vectors representing d_1-d_{12}

Summary

- Other models
  - Page rank model
  - Termsets
  - Fuzzy set model
  - Extended Boolean model
  - Generalised vector space