

Towards Context Modeling Using Vector Space Bases

Massimo Melucci
Department of Information Engineering
University of Padova
Italy
massimo.melucci@unipd.it

ESF/PESC Exploratory Workshop – IRiX
Glasgow, July 27th, 2005

ESF/PESC Exploratory Workshop – IRiX *Towards Context Modeling Using Vector Space Bases* Glasgow, July 27th, 2005

Background

- is information retrieval (IR) context-dependent?
 - IR deals with searching all and only the relevant documents for any information need and user
 - context is implied
 - relevance is context-dependent
 - interaction and feedback are implied
 - relevance feedback was an early try in system-centred models
 - studies date back to decades ago – e.g. Ingwersen, Blair

Outline

- Scenario
- Motivations
- Approach
- The Vector Space Model
- Modeling Context
- Context Change
- Relevance Feedback
- Conclusions and Work in Progress

Scenario

- searchers express information needs and authors implement documents by using descriptors
- context affects inter-relationships and meaning
 - after a series of queries about *computers*, the next query is likely not to include that key word
 - if location is close to Padua, the query include *Padua* and *Venice*
 - if a document is general, highly technical terms are unlikely to be included together with general terms
- different contexts, different descriptors, documents–queries mismatch

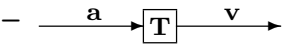
Motivations

- lack of representation of context in system-centred IR models
 - classical models were defined by assuming that there is one user, one information need for each query, one location, one time, one history, one profile
- ad-hoc techniques to capture time, space, histories, profiles are injected into models
 - sensors to get space location, log-files to implement history, metadata to describe profiles, clocks and calendar to get time

Approach

- adopt a classical IR model
 - before defining new models, let's see if the tradition suggests something
- the Vector Space Model
 - gives an intuitive view
 - proves effective for diverse media and languages
 - there are some yet-to-exploit potentialities, some examples:
 - * VSM re-evaluation (Wong and Raghavan)
 - * LSI (Dumais *et al.*)
 - * Geometry of IR (van Rijsbergen)

Definitions

- V is a vector space in \mathbb{R}^n , that is $V \subseteq \mathbb{R}^n$
- $\{\mathbf{t}_1, \dots, \mathbf{t}_m\}$ is a set of m column vectors in V :
 - $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_m]$ is the corresponding matrix
 - \mathbf{T} will be used as notation
- \mathbf{T} is linearly independent if $m \leq n$ and $\sum_{i=1}^m a_i \mathbf{t}_i = \mathbf{0}$ only if $a_i = 0$ for every i , i.e. no t can be linear combination of the t s
- if \mathbf{T} is linearly independent and $m = n$, \mathbf{T} is a *base* for V
- a base \mathbf{T} for V generates all the vectors of $\mathbf{v} \in V$
 - 

$$\mathbf{a} \rightarrow \boxed{\mathbf{T}} \rightarrow \mathbf{v}$$
 - $\mathbf{v} = \mathbf{T} \cdot \mathbf{a}$
 - there are more than one base

The Vector Space Model in Principle

- let $\{t_1, \dots, t_n\}$ by a set of unique descriptors
- two independent constructs:
 - the set of descriptors is modeled as a *base* $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_n]$
 - the *coefficients* a_1, \dots, a_n combine the base vectors to generate document or query vectors
- mathematically speaking,

$$\mathbf{v} = \sum_{i=1}^n a_i \mathbf{t}_i = \mathbf{T} \cdot \mathbf{a}$$

- ranking by the inner product between query and document vectors

$$\mathbf{d}^\top \cdot \mathbf{q} = \mathbf{a}^\top \cdot (\mathbf{T}^\top \cdot \mathbf{T}) \cdot \mathbf{b}$$

where $\mathbf{d} = \mathbf{T} \cdot \mathbf{a}$ and $\mathbf{q} = \mathbf{T} \cdot \mathbf{b}$

The Vector Space Model in Practice

- \mathbf{T} is orthogonal, i.e. $\mathbf{T}^\top \cdot \mathbf{T}$ is a diagonal matrix \mathbf{D}
- $\mathbf{d}^\top \cdot \mathbf{q} = \mathbf{a}^\top \cdot \mathbf{D} \cdot \mathbf{b}$
 - correlation are ignored, descriptor vectors are orthogonal
 - computation is much simpler
- \mathbf{D} is even the identity matrix, thus $\mathbf{d}^\top \cdot \mathbf{q} = \mathbf{a}^\top \cdot \mathbf{b}$
- \mathbf{T} is unique

Modeling Context

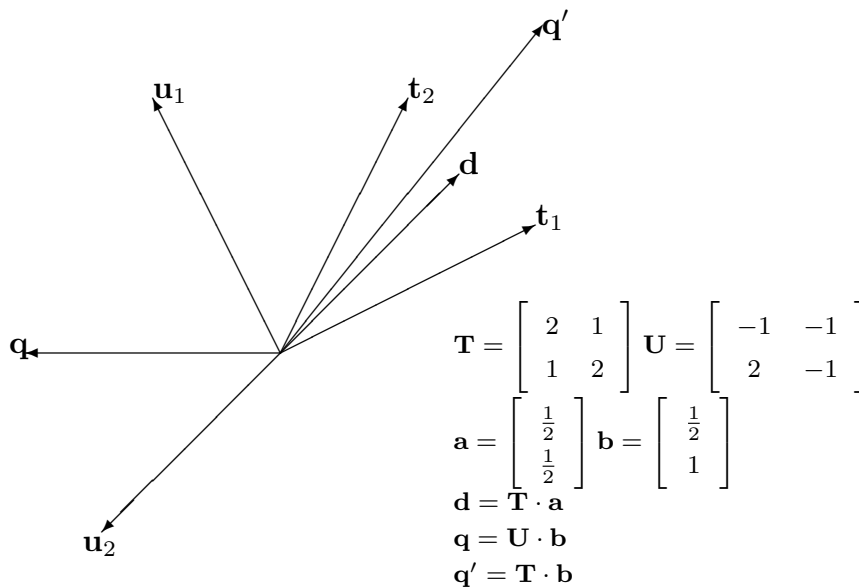
- linear independence is not assumed here
- \mathbf{T} is no longer unique, each document or query vector may be generated by its own base
- a *base* of a vector space is the construct to model *context*
 - the meaning of descriptors is given by the actual base vector components and correlation
 - context is modeled by base vector components and correlations
 - linear transformations between bases are matrices which model context changes

Modeling Context (cont.)

- the vector \mathbf{d} of the document d written in its own context is generated by \mathbf{T}
 - which is not necessarily equal to the base \mathbf{U} that generates, say, a query vector \mathbf{q} or to the base \mathbf{T}' that generates another document vector \mathbf{d}'
- if relevance is estimated by the usual inner product

$$\mathbf{d} = \mathbf{T} \cdot \mathbf{a} \quad \mathbf{q} = \mathbf{U} \cdot \mathbf{b} \quad \mathbf{d}^\top \cdot \mathbf{q} = \mathbf{a}^\top \cdot (\mathbf{T}^\top \cdot \mathbf{U}) \cdot \mathbf{b}$$

Modeling Context – example



Context Change

- context changes reflect on the descriptors used to describe document or query contents
- the descriptors describing documents or queries are represented as base vectors generating document or query vectors, respectively
- a change of the context represented by \mathbf{T} leads to a new context represented by \mathbf{U}

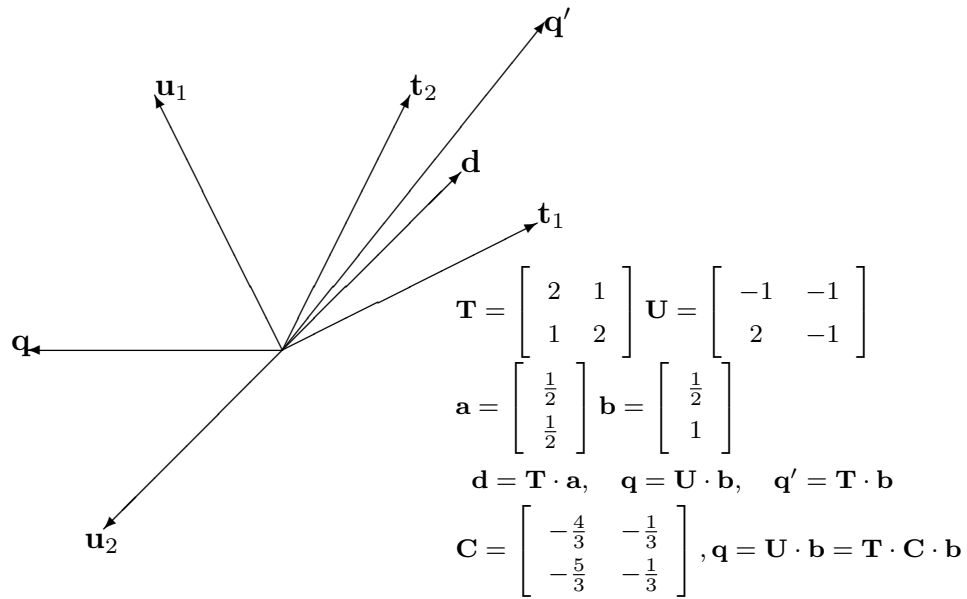
Context Change (cont.)

- for any \mathbf{T}, \mathbf{U} there exists a unique \mathbf{C} such that $\mathbf{U} = \mathbf{T} \cdot \mathbf{C}$
 - \mathbf{C} is the context change matrix which represents a linear transformation mapping \mathbf{T} to \mathbf{U}
 - one needs to only find the right matrix out to compute context change
- ranking:

$$\begin{aligned} \mathbf{d}^T \cdot \mathbf{q} &= \mathbf{a}^T \cdot (\mathbf{T}^T \cdot \mathbf{U}) \cdot \mathbf{b} \\ &= \mathbf{a}^T \cdot (\mathbf{T}^T \cdot \mathbf{T}) \cdot \mathbf{C} \cdot \mathbf{b} \\ &= \mathbf{a}^T \cdot (\mathbf{T}^T \cdot \mathbf{T}) \cdot \mathbf{c} \end{aligned}$$

- note that $\mathbf{c} = \mathbf{C} \cdot \mathbf{b}$ is the result of the rotation of \mathbf{b} to the context of \mathbf{d}

Modeling Context – example (cont.)



Relevance Feedback

- let q be a query and $\mathbf{q} = \mathbf{U} \cdot \mathbf{b}$ its generation by base \mathbf{U} unique for any query and document
- $0 < r < N$ relevant documents out of N documents

$$\begin{aligned} \mathbf{q}^+ &= \mathbf{q} + \frac{1}{r} \sum_{i=1}^r \mathbf{d}_i - \frac{1}{N-r} \sum_{j=r+1}^N \mathbf{d}_j \\ &= \sum_{k=1}^n b_k \mathbf{u}_k + \frac{1}{r} \sum_{i=1}^r \sum_{k=1}^n a_{ik} \mathbf{u}_k - \frac{1}{N-r} \sum_{j=1}^{N-r} \sum_{k=1}^n a_{jk} \mathbf{u}_k \\ &= \sum_{k=1}^n b_k \mathbf{u}_k + \sum_{k=1}^n \left(\sum_{i=1}^r \frac{a_{ik}}{r} \right) \mathbf{u}_k - \sum_{k=1}^n \left(\sum_{j=r+1}^N \frac{a_{jk}}{N-r} \right) \mathbf{u}_k \\ &= \sum_{k=1}^n \left(b_k + \sum_{i=1}^r \frac{a_{ik}}{r} - \sum_{j=r+1}^N \frac{a_{jk}}{N-r} \right) \mathbf{u}_k \\ &= \sum_{k=1}^n b_k^+ \mathbf{u}_k \end{aligned}$$

where $\mathbf{d}_h = \mathbf{U} \cdot \mathbf{a}_h = \sum_k a_{hk} \mathbf{u}_k$

Relevance Feedback (cont.)

- thus, $\mathbf{q}^+ = \mathbf{U} \cdot \mathbf{b}^+$ – recall that $\mathbf{q} = \mathbf{U} \cdot \mathbf{b}$
- there exists \mathbf{C} such that $\mathbf{b}^+ = \mathbf{C} \cdot \mathbf{b}$
- therefore,

$$\begin{aligned}\mathbf{q}^+ &= \mathbf{U} \cdot \mathbf{b}^+ \\ &= \mathbf{U} \cdot (\mathbf{C} \cdot \mathbf{b}) \\ &= (\mathbf{U} \cdot \mathbf{C}) \cdot \mathbf{b} \\ &= \mathbf{U}' \cdot \mathbf{b}\end{aligned}$$

where $\mathbf{U}' = \mathbf{U} \cdot \mathbf{C}$

- the context is provided by the partition of the collection in the set of relevant documents and its complement

Conclusions and Work in Progress

- can vector space bases model context?
- context discovery through descriptor correlations
- context exploitation through sensors
- conceptual issues in base vectors
- computational issues
- ...