Geometry and Information Retrieval

Alfonso E. Romero

aeromero@decsai.ugr.es

Dept. of Computer Science and Artificial Intelligence

Univ. de Granada. 18071 - Granada

1 Information Retrieval

With the arrival of the digital computer in the second half of the twentieth century, a vast amount of information has been stored and made available. The growing of accesible information has reached an exponential growing rate, and computer scientists have been worried about the problem of accessing and searching this information accurately.

The subfield of Computer Science that deals with the representation, automated storage and retrieval of information items is called information retrieval (IR) [10], [1], [12]. We denote these items as documents (unit of retrieval) which might be a paragraph, a section, a chapter, a web page, an article, or a whole book [1].

The two main views of an IR system are the following. The former, the indexing subsystem, which takes a set of documents and converts them to a suitable representation (what is called an index), and the latter (the most important one) retrieval subsystem which answers queries given from the user, with the subset of documents more relevant to that query.

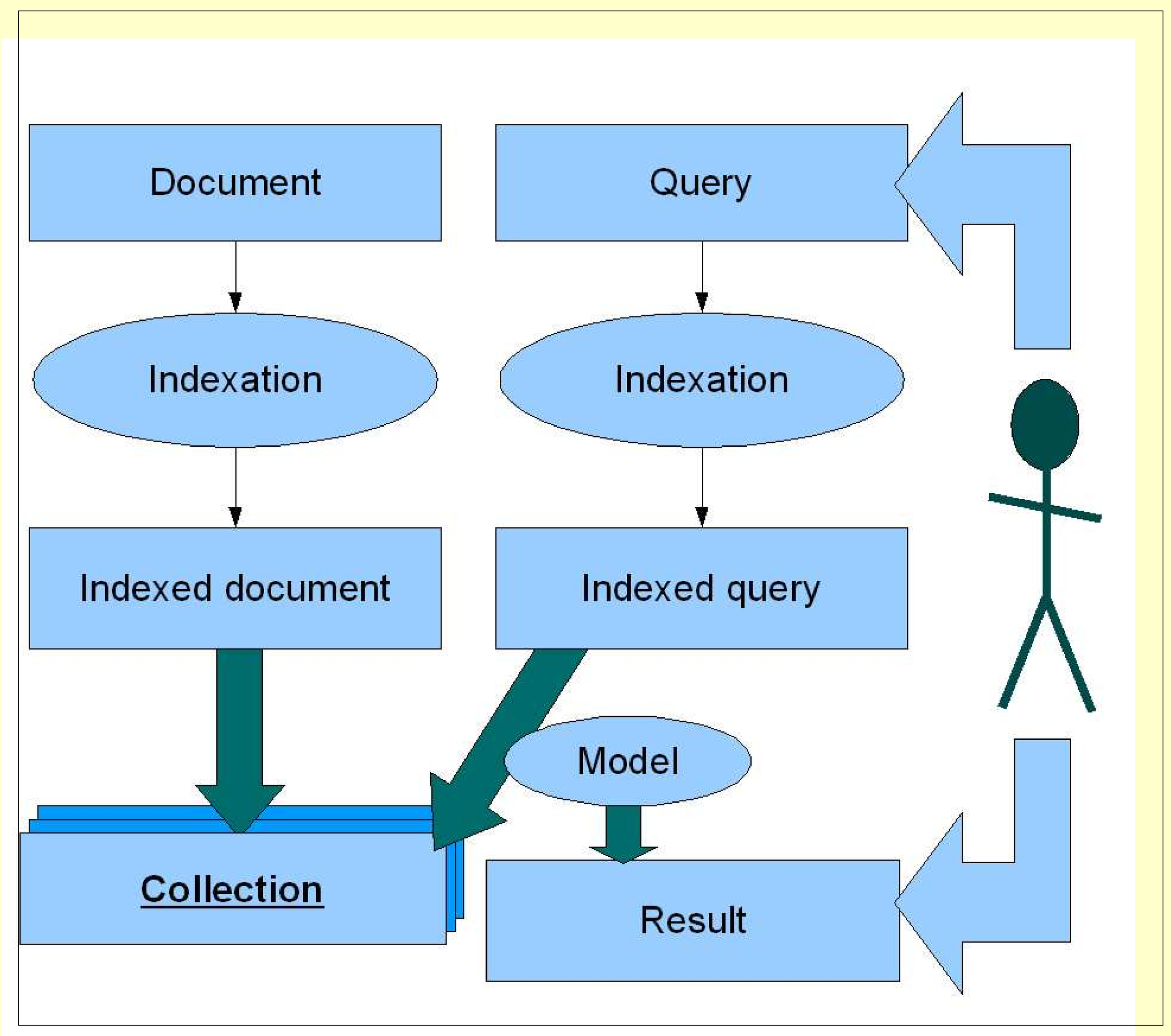

Figure 1: Conceptual view of an IR system.

While indexes are built using a set of well known data structures [2], [14], and all of them store almost the same information, the characteristical part of an IR system is the way it ranks documents, given a certain query. This is what is called the model.

The formal notion of an IR system is defined by Baeza [1] as a quadruple

$$\big(\mathcal{D},\ \mathcal{Q},\ \mathcal{F},\ \mathcal{R}(q_i, d_j)\big),$$

where:

- $\mathcal{D}$ is the set of the representations (logical views) of the documents.
- $\mathcal{Q}$ is the set of the representations of the queries.
- $\mathcal{F}$ is a framework or model to represent documents, queries and the relationships among them.
- R is a map ("ranking")

  $$R : \mathcal{D} \times \mathcal{Q} \to \mathbb{R}$$

  that associates a real number $R(q_i, d_j)$ to each query $q_i \in \mathcal{Q}$ and document $d_j \in \mathcal{D}$.

The most popular models in the literature [1] are based on boolean logic, linear algebra, probability or fuzzy set theory. As it is well known, there are many mathematical approaches to modeling these systems. We are interested here in some of them which use geometry at different levels.

2 The vector space model

The Vector Space Model (VSM) is one of the most successful models in IR. It was firstly proposed by G. Salton, A. Wong and C. S. Yang [13] in the 70s, and it is used until nowadays. It represents each document as an n-tuple of nonnegative real numbers. Each coordinate corresponds to one of the n different terms present in the collection. The value of the coordinate describes the importance of the term in the document, i. e., a very important term in the document has a higher value of the coordinate than less important ones. The value (weight) of the importance of the term i on the document j (the i-th component of the j-th document vector) is denoted by w_{ij}, and it is often defined as:

$$w_{ij} = \mathrm{tf}_{ij}\ \mathrm{idf}_i = \frac{f_{ij}}{\max_k f_{kj}} \log \frac{D}{D_i},$$

where f_{ij} is the absolute frequency of the i-th term in the j-th document. On the other hand, idf_i stands for inverse document frequency of i-th term, as the logarithm of the number of documents between the number of documents that term appears in, and it measures the rarity of the term in the collection.

An as interpretation, we can see a general form for the weighting scheme as following [4]:

$$w_{ij} = \text{local\_weight}_{ij} \cdot \text{global\_weight}_i \cdot \text{normalization}_j$$

A query is also represented as a vector of $\mathbb{R}^n$, usually having $q_i = 0$ if $t_i \notin q$ and $q_i = 1$ if $t_i \in q$. We will write w_{iq} for the weight of the i-th term on the query q.

Denote by $\langle\,,\rangle$ the usual scalar product of $\mathbb{R}^n$. The similarity between each document d_j and a query q is measured as the cosine of the angle of the two vectors, $\cos(\widehat{q, d_j})$:

$$R(q, d_j) = \frac{\langle q, d_j \rangle}{\|q\| \|d_j\|} = \frac{\sum_{k=1}^{n} w_{kq}\, w_{kj}}{\sqrt{\sum_{k=1}^{n} w_{kq}^2\ \sum_{k=1}^{n} w_{kj}^2}}$$

Both, queries and documents, live in the Euclidean metric vector space $\mathbb{R}^n$. Observe that it hold $w_{iq} \geq 0,\ q \neq 0$ and $w_{ij} \geq 0,\ d_j \neq 0$.

The ordering of the documents given by the VSM with the similarity function is equivalent to the ordering obtained using the intrinsic distance d over the unit sphere $\mathbb{S}^{n-1}$ [3, p. 279] where live the normalized vectors (queries and documents). In fact,

$$d(\tilde{q}, \tilde{d}) = \arccos R(q, d),$$

where $\tilde{q} = q/\|q\|$. Therefore, sorting documents by increasing similarity gives the same result than sorting normalized vectors by decreasing distance. More precisely, document and query vectors live in the part of the sphere with nonnegative coordinates, which is, topologically equivalent to the $(n-1)$-simplex.

A deeper analysis of the VSM and its variants is carried out in [9].

3 Standard quantum logic

Let H be an $n(>2)$-dimensional vector metric space, and let $\mathcal{P}$ be the set of all orthogonal projectors of H. Given a subspace E of H, we denote by p_E the orthogonal projector on E. We can define, for any $q, r \in \mathcal{P}$ the following operators:

$$\bar{q} = 1 - q, \qquad q \wedge r = p_{q(H) \cap r(H)}, \qquad q \vee r = p_{q(H) + r(H)}$$

This logic, called standard quantum logic, was introduced, in a more general context of Hilbert spaces, by Birkhoff and Von Neumann [5]. It should be observed that it is not a "classical logic". In fact, distributive laws do not hold on $\mathcal{P}$ in general (instead, only a weaker law, "modularity" is satisfied here). Therefore, $\mathcal{P}$ is not a boolean algebra but only a lattice (more precisely an orthomodular lattice [8]).

4 Probability measures of subspaces

The aim here is to define a probability measure over the set of subspaces of H, using the natural identification of each subspace with the corresponding orthogonal projector. A probability measure on subspaces of $\mathbb{R}^n$ is a map

$$\mu : \{L : L \text{ is a subspace of } \mathbb{R}^n\} \longrightarrow \mathbb{R},$$

which satisfies the following properties:

1. $\mu(\{0\}) = 0$.
2. $\mu(\mathbb{R}^n) = 1$.
3. If L_i and L_j are subspaces of $\mathbb{R}^n$ such that $L_i \cap L_j = \{0\}$, then $\mu(L_i + L_j) = \mu(L_i) + \mu(L_j)$.

It is easily seen that for each positive definite self-adjoint operator $\mathbf{T}$ of $\mathbb{R}^n$, with $\mathrm{trace}(\mathbf{T}) = 1$, a probability measure $\mu_{\mathbf{T}}$ can be defined by setting

$$\mu_{\mathbf{T}}(L) := \mathrm{trace}(\mathbf{T} \circ \mathbf{P}_L),$$

where $\mathbf{P}_L : \mathbb{R}^n \to \mathbb{R}^n$ is the orthogonal projection on L.

Conversely, in a much more general setting, it was proved [6] that every such probability measure is constructed in this way.

In [11], C. J. van Rijsbergen relates the VSM, probability measures of subspaces and probabilistic IR models [7] as follows.

Given a query q and a document d, denote by $\mathbf{T}_{\tilde{q}}, \tilde{q} = q/\|q\|$, the operator defined, with respect to the standard basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$ by

$$\mathbf{T}_{\tilde{q}}(e_j) = \sum_{i=1}^{n} (\tilde{q}_i\, \tilde{q}_j)\ e_i,$$

and by $\mathbf{P}_d (= \mathbf{P}_{\tilde{d}})$ the orthogonal projection on the 1-dimensional subspace spanned by d. Then, using the probability measure $\mu_{\mathbf{T}_{\tilde{q}}}$, we have:

$$\mu_{\mathbf{T}_{\tilde{q}}}(\,\text{Span}(\{d\})\,) = \mathrm{trace}(\mathbf{T}_{\tilde{q}} \circ \mathbf{P}_d) = \langle \tilde{d}, \tilde{q} \rangle^2 = \cos^2(\widehat{\tilde{d}, \tilde{q}})$$

This result links the probability measures of subspaces to the VSM, and it can be also interpreted as the probability of relevance of the document d, given the query q.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, **Modern Information Retrieval**, Addison Wesley Longman, 2002.

[2] R. Baeza-Yates and W. B. Frakes (eds.), **Information Retrieval. Data Structures and Algorithms**, Prentice Hall, 1992.

[3] M. Berger, **Geometry II**, Universitext, Springer-Verlag, 1987.

[4] M. W. Berry and M. Browne, **Understanding Search Engines - Mathematical Modeling and Text Retrieval**. SIAM, Philadelphia, 2000.

[5] G. Birkhoff and J. Von Neumann, *The Logic of Quantum Mechanics*, Ann. Math., **37**(4), (1936), 823–843.

[6] R. Cooke, M. Keane and W. Moran, *An elementary proof of Gleason's Theorem*, Math. Proc. Camb. Phil. Soc., **98**, (1985), 117–128.

[7] F. Crestani, M. Lalmas, C. J. van Rijsbergen and I. Campbell, *"Is This Document Relevant? ... Probably": A Survey of Probabilistic Models in Information Retrieval*, ACM Comput. Surv., **30**(4) (1998), 528–552.

[8] G. Kalmbach, **Orthomodular Lattices**, Academic Press, London, 1983.

[9] V. V. Raghavan and S. K. M. Wong, *A Critical Analysis of Vector Space Model for Information Retrieval*, J. Am. Soc. Inform. Sci., **37**(2), (1986), 279–287.

[10] C. J. van Rijsbergen, **Information Retrieval**, (Second edition), Butter Worths, London, 1979.

[11] C. J. van Rijsbergen, **The Geometry of Information Retrieval**, Cambridge University Press, 2004.

[12] G. Salton and M. J. McGill, **Information Retrieval**, MacGraw-Hill, 1987.

[13] G. Salton, A. Wong and C. S. Yang, *A Vector Space Model for Automatic Indexing*, Commun. ACM, **18**(11), (1975), 613-620.

[14] I. H. Witten, A. Moffat and T. C. Bell, **Managing Gigabytes**, **Morgan Kaufmann**, 1999.