# Comparaison du modèle vectoriel et de la pondération tf*idf associée avec une méthode de propagation d'activation

**Christophe Brouard**

*UPMF-Grenoble2/CNRS*
*LIG UMR 5217/équipe AMA*
*Grenoble, France*
*Christophe.Brouard@imag.fr*

RÉSUMÉ. *L'objet de ce papier est de montrer qu'il est possible de mettre le modèle vectoriel et la pondération tf*idf associée en correspondance avec le calcul d'une résonance dans un réseau associatif basé sur une méthode très simple de propagation d'activation. Nous décrivons un réseau associatif associant termes et documents puis un calcul de résonance entre une requête et un document dans ce réseau. La pondération tf*idf apparaît naturellement dans le calcul et nous expliquons comment chacun des facteurs s'interprète dans la propagation d'activation. Nous montrons que ce calcul, comme le cosinus, correspond au produit de l'évaluation de la spécificité et de l'exhaustivité du document vis-à-vis de la requête. La comparaison expérimentale réalisée sur les corpus CLEF3 et TREC3 révèle que, si l'on choisit adéquatement les fonctions qui régissent l'activation et la propagation d'activation, la méthode basée sur la résonance obtient des performances similaires à celles du modèle Okapi-BM25.*

ABSTRACT. *We show in this paper that it is possible to establish a correspondence between the vector space model and a spreading activation method based on a resonance computation in a network linking the terms and the documents. We describe the network and the resonance computation between a query and a document in the network. The tf*idf weighting scheme is naturally introduced in the computation and we explain how these factors can be interpreted in the spreading activation method. Then, we show that, like the cosine in the vector space model, the resonance computation corresponds to the product between the measures of specifity and exhaustivity of the document for the query. The experiments carried out on the CLEF3 and TREC3 datasets show that the performances of the model based on resonance are similar to those of the Okapi-BM25 model.*

MOTS-CLÉS: *modèle de RI, fonction de correspondance, méthode de propagation d'activation*

KEYWORDS: *IR model, scoring function, spreading activation method*

## 1. Introduction

La définition d'un modèle de Recherche d'Information (RI) comprend la définition d'une représentation de la requête et du document et la définition d'une fonction de correspondance qui calcule le score des documents pour les différentes requêtes à partir de leurs représentations respectives. La fonction de correspondance correspond donc au cœur de tout système de RI et de nombreux modèles, qui se distinguent par les fonctions de correspondance qu'ils ont adoptées ont été proposés. Le modèle vectoriel initié par Salton (1975) et développé par la suite pour arriver à la forme qu'on lui connaît actuellement (Manning et al, 2008) propose par exemple de calculer le cosinus de l'angle formé par les vecteurs représentant le document et la requête. Certains modèles s'attèlent au calcul de la probabilité de pertinence sachant un document et une requête (Robertson et Jones, 1976). D'autres encore mesurent la probabilité que la requête soit issue du modèle de langue que constitue un document (Ponte et Croft, 1998). Sans chercher à être exhaustif on peut aussi mentionner les modèles qui reposent sur une mesure de l'écart entre la fréquence des termes de la requête dans le document et la fréquence de ces mêmes termes dans la collection (Amati et Van Rijsbergen, 2002).

La fonction de correspondance est directement liée à la notion de pertinence et sa formalisation. Pourtant, tous les modèles ne s'appuient pas sur une approche claire de la notion de pertinence. La justification de la fonction de correspondance est plus souvent liée aux intuitions concernant l'impact de différents facteurs dans un calcul et aux résultats expérimentaux qu'à une réflexion théorique sur la notion de pertinence. Le modèle que nous proposons s'appuie sur la notion de résonance introduite par Grossberg (1976) dans le contexte des réseaux de neurones. Ce concept s'accorde bien avec les notions de spécificité et d'exhaustivité souvent mentionnées dans les études menées sur la notion de pertinence (Brouard, 2004). Il a une plausibilité neurophysiologique et une signification très générale concernant l'interaction d'un système avec son environnement. Nous proposons donc d'écrire la pertinence d'un document vis-à-vis d'une requête comme une résonance dans un réseau de neurones très simple où termes et documents sont associés. La formalisation fait apparaître une fonction de correspondance très proche de celle du modèle vectoriel.

La deuxième partie de cet article établit un lien entre notre proposition et d'autres travaux de recherche. La troisième partie décrit le réseau dans lequel est calculée la résonance et le calcul de résonance proprement dit. Dans la quatrième partie, nous précisons notre modèle en choisissant adéquatement les fonctions régissant l'activation et la propagation d'activation et nous le comparons avec la forme la plus usuelle du modèle vectoriel. Nous montrons de plus que comme la plupart des modèles de RI, les deux facteurs correspondant à la spécificité et l'exhaustivité du document vis-à-vis de la requête sont au cœur de notre modèle. Nous montrons plus spécifiquement que comme le cosinus utilisé dans le modèle vectoriel, le calcul de résonance correspond au produit des 2 facteurs. Dans la cinquième partie, nous

vérifions la qualité du modèle sur les corpus TREC3 et CLEF3. Dans la dernière partie, nous concluons en dressant un bilan du travail et en discutant de l'universalité de la méthode et de son application à d'autres problèmes de sélection d'information.

## 2. Travaux reliés

Différents travaux sur la notion de pertinence ont été menés. Néanmoins, rares sont les travaux qui permettent de faire un lien entre la notion de pertinence et une fonction de correspondance. On peut néanmoins citer les travaux de Van Rijsbergen (1986), Wilson (1973), Devlin (1991) et Nie (1988). Dans ces travaux, la notion d'implication est mise en exergue. Dans le cas de la RI elle se définit entre la requête et le document. L'implication d→q correspond à l'inclusion des termes de la requête dans ceux du document. Elle rend compte du fait que le document aborde tous les points de la requête (on parle d'exhaustivité). L'implication inverse q→d qui correspond à l'inclusion des termes de la requête dans ceux du document rend compte du fait que le document ne traite que des points abordés dans la requête et pas d'autre chose (on parle de spécificité). Ces deux implications se retrouvent implicitement dans les différents modèles bien qu'elles ne les fondent pas. On peut ainsi émettre l'hypothèse comme Nie (1988), que la pertinence est une combinaison de la spécificité et de l'exhaustivité. De plus, il est possible d'établir une analogie entre ces notions et des mécanismes de propagation d'activation dans des réseaux de neurones et plus particulièrement la notion de résonance (Brouard, 2004). Cette analogie donne une plausibilité neurophysiologique à cette formalisation de la pertinence. Par ailleurs, Grossberg (1976) qui a introduit cette notion dans sa théorie de la résonance adaptative (ART) décrit le mécanisme de résonance comme un mécanisme de sélection d'information garantissant pour un système plongé dans un environnement un compromis entre adaptativité et stabilité. D'un point de vue plus général, cette formalisation de la notion de pertinence a donc aussi une signification en terme d'évolution d'un système dans son environnement (Grossberg, 1999).

Enfin, l'idée de sélection d'information par propagation dans un réseau (neuronal ou sémantique) n'est pas nouvelle. L'idée consiste à activer dans le réseau des neurones ou concepts représentant une situation (objectif, environnement, etc...) et à laisser se propager l'activation aux informations liées dans le réseau. Les premières méthodes de propagation d'activation ont été définies dans le domaine de la Psychologie Cognitive pour modéliser l'accès aux connaissances en mémoire (Collins et Loftus, 1975). Anderson (1983) a utilisé ce type de méthodes comme mécanisme général de sélection de connaissances pour proposer un modèle général de la cognition (ACT). De nombreuses tentatives ont aussi été faites en Recherche d'Information (Crestani, 1997). Néanmoins, les différentes méthodes définies se sont heurtées au problème du contrôle de la propagation. En effet, si aucune règle n'est introduite, l'activation tend à se propager à tout le réseau ce qui va totalement à l'encontre de l'objectif d'une méthode de sélection. L'ajout de contraintes qui permettent de contrôler l'activation de façon efficace est possible

(Cohen et Kjeldsen, 1987) mais dans ce cas on perd un peu l'intérêt de la méthode, l'intelligence n'étant plus dans la propagation d'activation mais dans un système de règles de contraintes qui ne s'appliquent de plus qu'à un problème particulier et les méthodes de propagation semblent alors un peu artificielles. Nous pensons que l'idée de résonance permet de contrôler "naturellement" la propagation et est suffisamment générale pour être appliquée à de nombreux problèmes différents.

## 3. Description du modèle

### 3.1. *Description du réseau et de sa construction (indexation)*

Le réseau considéré est très simple (Figure 1). Il s'agit d'un réseau à 2 couches. Dans la première couche, chaque nœud représente un terme. Dans la seconde couche, chaque nœud représente un document. Des connections n'existent qu'entre les nœuds de couches différentes. Une connection entre un document *d* et un terme *t* n'existe que si le document contient le terme. Le poids de la connection $w_{td}$ entre *t* et *d* est croissant avec le nombre d'occurrences du terme dans le document (sa valeur sera précisée par la suite) et décroissant avec le nombre de termes dans le document.

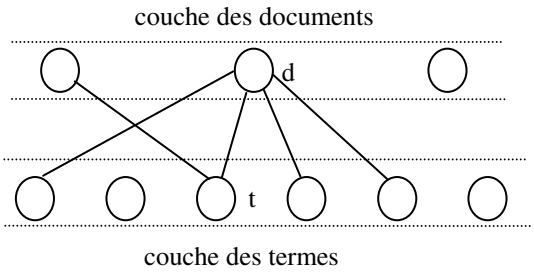


**Figure 1.** *Le réseau est composé d'une couche de termes et une couche de documents. Une connection entre un terme t et un document d existe si d contient t.*

La prise en compte des documents est incrémentale. Lorsqu'un nouveau document est pris en compte, les nœuds correspondant aux termes du document sont créés s'ils n'existaient pas déjà. Le nœud correspondant au document est créé et les connections liant les termes et le document sont créées et valuées. Nous ne le détaillerons pas ici mais on peut parfaitement représenter le mécanisme de construction par l'application d'une simple règle de Hebb (renforcement de la connection de nœuds activés simultanément). La prise en compte d'un document correspond alors simplement à activer les nœuds correspondant aux termes présents dans le document et le nœud correspondant au document. Si on considère l'activation du document totale et égale à 1, le poids de la connection résultant du produit des activations (règle de Hebb) correspond donc à l'activation du terme.

### 3.2. *Description du calcul de résonance (recherche)*

Une requête est représentée par un pattern d'activation dans la couche des termes. L'activation *a(t)* d'un terme *t* de la requête *q* dépend du nombre d'occurrences du terme dans la requête. Le calcul de la résonance *R(d,q)* d'un document *d* pour une requête *q* correspond à la mesure de la quantité d'activation conservée lors de la propagation de *q* à *d* puis de *d* à *q* puis de *q* à *d*, etc...

Pour calculer cette quantité, on considère une méthode de propagation régie par 3 règles. La première règle est que l'activation d'un nœud "document" correspond simplement à la somme des activations reçues. La deuxième est l'activation d'un nœud "terme" correspond au produit de la somme des activations reçues par son activation initiale $a_0(t)$. La troisième règle est que l'activation transmise par un nœud $n_i$ à un nœud $n_j$ est le produit de 3 facteurs :

- l'activation du nœud $n_i$ : $a(n_i)$
- le poids de la connection entre les nœuds $n_i$ et $n_j$ : $w_{ij}$
- un facteur de division : $div(n_i)$ qui décroît avec le nombre de connections du nœud (l'activation du nœud se répartit dans une certaine mesure sur ses connections).

En considérant les règles de propagation énoncées précédemment, l'activation du document *d* après propagation de l'activation de la requête *q* correspond à :

$$a_1(d) = \sum_{t \in d \cap q} a_0(t).w_{td}.div(t) \qquad [1]$$

L'activation d'un terme *t* après rétro-propagation de *d* vers la couche des termes est :

$$a_1(t) = a_1(d).w_{td}.div(d).a_0(t) \qquad [2]$$

L'activation de *d* après une nouvelle propagation de la couche des termes vers *d* est :

$$a_2(d) = \sum_{t \in d \cap q} a_1(t).w_{td}.div(t) \qquad [3]$$

En écrivant $a_n(d)$ en fonction de $a_{n-1}(d)$ on trouve que l'activation du document *d* après *n* étapes de propagation s'écrit pour n≥1 comme une suite géométrique :

$$a_n(d) = \left[ \sum_{t \in d \cap q} a_0(t).w_{td}^{\,2}.div(t).div(d) \right]^{n-1}.a_1(d) \qquad [4]$$

Si la raison de la suite est inférieure à 1, la suite converge vers 0. Sinon elle diverge. Quoi qu'il en soit, lorsque *n* grandit, le facteur prépondérant pour l'ordre des documents est la raison de la suite que l'on considèrera comme la mesure de résonance *R(d,q)* du document *d* vis-à-vis de la requête *q* :

$$R(d,q) = \sum_{t \in d \cap q} a_0(t).{w_{td}}^2.div(t).div(d) = div(d).\sum_{t \in d \cap q} a_0(t).{w_{td}}^2.div(t) \qquad [5]$$

## 4. Instanciation du modèle et comparaison avec le modèle vectoriel

### 4.1. Poids des termes dans le document et la requête

Traditionnellement en Recherche d'Information, le poids d'un terme dans un document correspond au facteur *tf* (term frequency) qui correspond au nombre d'occurrences du terme dans le document. On peut aussi considérer des formes plus complexes tenant compte de la taille du document tel que le *tf* qui apparaît dans Okapi-BM25 (Robertson et Zaragoza, 2009). Dans nos expérimentations, nous avons considéré un *tf* prenant comme Okapi, la forme générale *(1/1+(1/x))*. Dans notre processus d'indexation cette fonction peut être interprétée comme la fonction d'activation des termes. A la différence du *tf* d'*Okapi-BM25*, nous considérons un poids compris dans l'intervalle [0,1] :

$$f(t,d) = \frac{1}{1 + (c_1 + c_2 * dl) / tfb} \qquad [6]$$

où *dl* correspond au nombre de termes dans le document, *tfb* correspond au nombre d'occurrences du terme dans le document et où $c_1$ et $c_2$ sont des constantes. Dans le but de rendre la comparaison avec le modèle vectoriel aisée (comme on le verra par la suite) on choisit de définir le poids $w_{td}$ comme la racine carrée de *f(t,d)*. Concernant la représentation de la requête et donc l'activation initiale des termes, on considèrera *f(t,q)*, le nombre d'occurrences des termes dans la requête.

### 4.2. Fonction de répartition de l'activation sur les connections du nœud

La fonction *div* doit être décroissante avec le nombre de connections du nœud. Elle correspond à la notion de "fan-out" évoquée dans les méthodes de propagation (Crestani, 1997). Un choix intéressant puisqu'il nous ramènera à des notions connues consiste à utiliser la forme suivante pour calculer le facteur de division du nœud *n* :

$$div(n) = \log\left(\frac{nombre\ de\ connections\ potentielles}{nombre\ de\ connections\ de\ n}\right) \qquad [7]$$

Dans le cas d'un nœud "terme", le nombre de connections correspond au nombre de documents contenant ce terme ce qui correspond à ce qui est communément noté *df* (pour document frequency). Toujours dans le cas d'un nœud "terme" le nombre de connections potentielles correspond au nombre de documents. On obtient ainsi pour les nœuds "termes" le traditionnel facteur *idf* (pour inverse document frequency) que l'on notera *idf(t)*. Pour les nœuds "documents", on obtient le log d'un rapport entre le nombre de termes dans la collection et le nombre de termes dans le document que l'on notera *idf(d)*.

### 4.3. Comparaison du modèle après instanciation

En remplaçant dans l'équation 5 les différentes fonctions générales par leurs instanciations, on obtient :

$$R(d,q) = idf(d).\sum_{t \in d \cap q} f(t,q).f(t,d).idf(t) \qquad [8]$$

Cette formule correspond à la fonction de correspondance du modèle vectoriel dans laquelle, pour tenir compte de la taille du document, on a remplacé la norme du document par *idf(d).* Dans le cas du modèle vectoriel, la pondération *tf***idf* est intégrée au calcul du cosinus sans véritable justification. Dans le cas de notre méthode de propagation d'activation, elle apparaît naturellement.

La proximité des 2 modèles n'est pas totalement étonnante dans la mesure où dans le cas du cosinus comme dans le cas du calcul de résonance on calcule le produit d'une mesure d'exhaustivité et d'une mesure de spécificité. En effet, si l'on fait abstraction du poids des termes (on les suppose tous égaux à 1), le cosinus entre le vecteur requête et le vecteur document s'exprime de la façon suivante :

$$\cos(d,q) = \frac{|d \cap q|}{\sqrt{|d|}\sqrt{|q|}} \qquad [9]$$

où |X| désigne la cardinalité de l'ensemble X. On peut prendre le carré sans changer l'ordre des documents pour faire apparaître 2 facteurs mesurant respectivement l'inclusion des termes de *d* dans ceux de *q* (spécificité) et des termes *q* dans ceux de *d* (exhaustivité) :

$$\cos(d,q)^2 = \frac{|d \cap q|}{|d|}\frac{|d \cap q|}{|q|} \qquad [10]$$

Dans le cas de la méthode de propagation d'activation, la propagation de *q* vers *d* mesure l'inclusion des termes de *q* dans ceux de *d*. En effet, plus le nombre de termes de *q* présents dans *d* est important plus la quantité d'activation transmise est

importante (on fait la somme). Dans le sens inverse, la propagation de $d$ vers $q$ mesure l'inclusion des termes du document dans ceux de la requête. En effet, la division de l'activation sur les nœuds connectés a pour conséquence de conditionner la transmission de la totalité de l'activation du document à la requête au fait que tous les termes du document sont inclus dans l'ensemble des termes de la requête. Moins l'inclusion est forte plus la quantité d'activation perdue est grande.

## 5. Expérimentations

### 5.1. *Corpus utilisés*

Les corpus TREC3 et CLEF3 sont des corpus provenant de campagnes d'évaluation (respectivement TREC[1] et CLEF[2]). TREC3 est composée d'environ 740 000 documents et 100 requêtes. CLEF3 est composé d'environ 170 000 documents et 60 requêtes. Pour le corpus TREC3, seul le titre de la requête est utilisé. Pour le corpus CLEF3, le titre et la description sont utilisés. Les expérimentations ont été réalisées avec la bibliothèque Terrier-3.5[3] et les habituelles opérations de prétraitements (suppression des mots vides et extraction des stems avec l'algorithme de Porter) ont été réalisées.

### 5.2. *Systèmes comparés*

Comme nous l'avons expliqué précédemment notre modèle de propagation d'activation nous a permis de retrouver la forme générale du modèle vectoriel. Néanmoins, notre calcul nous a amené à introduire de petites différences par rapport à la version la plus usuelle du modèle vectoriel. Ces différences concernent la prise en compte de la taille du document (*idf(d)* à la place de la norme de *d*) et la forme du *tf*. Afin de vérifier la qualité des performances, nous avons donc comparé notre système (RESO) au système Okapi-BM25 (Robertson et Zaragoza, 2009) dont le niveau de performance est en général très bon et dont la forme du *tf* est similaire à celle que nous avons considérée. Les paramètres sont $b$, $k\_1$ pour BM25 et $c_1$ et $c_2$ pour RESO. Nous avons divisé l'ensemble de requêtes en deux et nous avons suivi un processus de validation croisée. Ainsi, la première moitié des requêtes a d'abord été utilisée pour le réglage des paramètres (en optimisant la MAP) et l'autre moitié pour les tests puis nous avons ensuite procédé dans l'ordre inverse. Nous avons donc obtenu 2 résultats par corpus (un par ensemble de requêtes) que nous avons ensuite moyennés.

1. http://terrier.org/
2. http://trec.nist.gov
3. http://www.clef-initiative.eu/

### 5.3. *Résultats et discussion*

Les résultats obtenus (table 1) montrent que l'on peut obtenir avec notre système des résultats similaires à ceux obtenus avec BM25. La différence est à l'avantage de BM25 mais elle est assez faible. Elle est respectivement de 0.4 et 0.3 de MAP pour CLEF3 et TREC3. Le niveau de performance atteint permet donc de conclure à l'efficacité de notre variante du modèle vectoriel.

| | **CLEF3** | **TREC3** |
|---|---|---|
| **RESO** | 45.0 - 41.5 - 33.7 - 18.0 | 25.6 - 51.6 - 52.2 - 41.3 |
| **BM25** | 45.4 - 42.6 - 34.1 - 18.0 | 25.9 - 54.4 - 53.0 - 42.2 |

**Table 1.** *Comparaison des performances de BM25 et RESO. Le premier nombre correspond à la MAP pour les 10 000 premiers documents retrouvés. Les 3 nombres suivants correspondent respectivement à la précision à 5, 10 et 50 documents.*

## 6. Conclusion et perspectives

Nous avons montré que le calcul d'une résonance dans un réseau associant termes et documents permettait de retrouver la forme générale de la fonction de correspondance du modèle vectoriel. Nous avons aussi montré que dans ce modèle la pondération tf*idf s'introduisait naturellement. Le choix des fonctions instanciant $w_{td}$ et *div* n'a certes pas été fait au hasard. Notre choix était guidé par l'objectif de se rapprocher du modèle vectoriel. Néanmoins les fonctions choisies sont cohérentes avec les idées d'association et de répartition de l'activation qui sous-tendent la méthode de propagation. Par ailleurs, nous avons montré que comme le cosinus notre calcul correspondait à un produit entre la spécificité et l'exhaustivité. La proximité entre notre modèle et le modèle vectoriel n'est donc pas seulement le fruit d'un simple jeu de réécriture.

Ce modèle de RI présente plusieurs intérêts. D'une part il s'appuie sur une formalisation théorique de la notion de pertinence. D'autre part, il a une plausibilité neurophysiologique puisqu'il ne s'appuie que sur des mécanismes neuromimétiques (association, propagation). Nous pensons que ce mécanisme de sélection peut être appliqué à de nombreux autres problèmes de sélection d'information. Il a par exemple été appliqué avec succès au problème de la classification de documents (Brouard, 2012) où la sélection porte sur la classe du document. Il pourrait l'être aussi dans la tâche d'extension de requête où le problème consisterait à sélectionner des termes à ajouter à la requête. Il pourrait l'être enfin aussi dans le cas de la désambiguïsation de mot où le problème consiste à sélectionner le bon sens d'un mot qui en possède plusieurs sur la base de son contexte. Plusieurs autres applications de la méthode sont en cours.

## 7. Bibliographie


Amati G., Van Rijsbergen C.J, Probabilistic models of information retrieval based on measuring the divergence from randomness, *ACM Trans. Inf. Syst*., 20(4), 357–389, 2002.

Anderson J.R., *The Architecture of Cognition*, Cambridge, MA Harvard Univ. Press, 1983.

Anderson J.R., *A Spreading Activation Theory of Memory*, Cambridge, *Journal of Verbal Learning and Verbal Behavior*, 22, 261-295, 1983.

Brouard C., Nie J.Y., Relevance as Resonance: a New Theoretical Perspective and a Practical Utilization in Information Filtering, *Inform. Process. and Management*, 40, 1-19, 2004.

Brouard C., Document Classification by Computing an Echo in a Very Simple Neural Network, *ICTAI*, 735-741, 2012.

Cohen P.R., Kjeldsen R., Information Retrieval by Constrained Spreading Activation in Semantic Networks, *Information Processing and Management*, 23(4), 255-268, 1987.

Collins A.M., Loftus E.M., A Spreading Activation Theory of Semantic Processing. *Psychological Review*, 82, 407-428, 1975.

Crestani F., Application of Spreading Activation Techniques in Information Retrieval, *Artificial Intelligence Review*, 11, 453-498, 1997.

Devlin K., *Logic and Information*. Cambridge University Press. 1991.

Grossberg S., Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors, *Biol. Cyber*., vol. 23, 117-140, 1976.

Grossberg S., The Link between Brain Learning Attention and Consciousness*, Consciousness and Cognition*, 8, 1-44, 1999.

Nie J-Y., An Outline of a General Model for Information Retrieval. *ACM-SIGIR,* 495-506, 1988.

Ponte J.M., Croft W.B. A Language Modeling Approach to Information Retrieval. *Research and Development in Information Retrieval*, *ACM-SIGIR*, 275-281, 1998.

Robertson S.E., Jones K.S., Relevance Weighting of Search Terms, *Journal of the American Society for Information Science*, 27(3), 129–146, 1976.

Robertson S.E., Zaragoza H., The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval, 3(4), 333-389, 2009.

Manning C., Raghavan D., Prabhakar; Schütze H., Introduction to Information Retrieval, Cambridge University Press. 2008.

Salton G., Wong A., Yang C.S., A vector space model for information retrieval, *Communications of the ACM*, 18(11), 613–620, 1975.

Van Rijsbergen C.J., A non-classical logic for information retrieval, *The Computer Journal*, 29(6), 481-485, 1986.

Wilson P., Situational Relevance. *Information Storage and Retrieval*, 9(8), 457-471, 1973.