INFORSID 2005

Evaluation : nouvelle approche avec juges

Didier NAKACHE [(1) et (2)], Elisabeth METAIS [(1)]

datamining@wanadoo.fr, metais@cnam.fr

(1) CEDRIC /CNAM : 292 rue Saint Martin - 75003 Paris, France
(2) CRAMIF : 17 / 19 rue de Flandre - 75019 Paris, France

MOTS CLES : évaluation, mesure, classification, catégorisation, TALN, F-mesure

KEYWORDS: evaluation, measure, classification, categorization, NLP, F-measure.

RÉSUMÉ : *la F-mesure est l'indicateur de synthèse communément utilisé depuis 25 ans pour évaluer les algorithmes de classification de données textuelles, à partir de la précision et du rappel. Elle est utilisée indifféremment pour les classifications et les catégorisations. Pour les problématiques de recherche d'information et de classification, certains préfèrent utiliser le Break Even Point. Cependant, ces mesures présentent quelques défauts : elles répondent à une logique binaire et ne permettent pas d'intégrer de niveaux d'exigences en matière de résultats, tant sur la pertinence que sur l'ordre de présentation des résultats.*

Ce papier propose une nouvelle approche de l'évaluation. Dans un premier temps, nous proposons de distinguer la classification de la catégorisation sur le plan sémantique pour aboutir à des propositions méthodologiques adaptées à chacune de ces situations. Puis, nous introduisons une nouvelle mesure : la K-mesure, applicable à ces deux problématiques, sur ensemble de la F-mesure et du Break Even Point. De plus, elle permet de prendre en compte des niveaux d'exigences. Il s'agit donc d'une méta mesure de convergence entre ces différentes approches, dont elle conserve les propriétés majeures. Enfin, nous proposons une méthodologie adaptée au problème spécifique de la recherche d'informations.

ABSTRACT. F-measure is an indicator used since 25 years to evaluate classification algorithms in textmining, from precision and recall. For classification and information retrieval, some ones prefer to use the break even point. Nevertheless, these measures have some inconvenient: they use a binary logic and don't allow applying a user (judge) assessment. This paper proposes a new approach of evaluation. First, we distinguish classification and categorization from a semantic point of view. Then, we introduce a new measure: the K-measure, which is an overall of F-measure and break even point, and allows applying user requirements. Finally, we propose a methodology for evaluation.

1. Introduction

Le traitement du langage naturel donne naissance à de nombreux algorithmes de recherche d'informations, classification, ou catégorisation. L'efficacité de ces algorithmes est calculée à partir de diverses mesures, dont les plus connues sont la précision et le rappel. Pour faciliter la lecture de la performance, [Van Rijsbergen 79] a créé une mesure synthétique : la F-mesure, qui combine ces deux indicateurs. Aujourd'hui, les besoins se sont diversifiés, les problématiques se sont complexifiées, mais l'on conserve les mêmes indicateurs de référence depuis 25 ans [Sparck Jones 2001]. Certes, cela facilite la comparaison entre différentes méthodes mais leur emploi est il toujours pertinent ? Comment intégrer de nouvelles considérations sans renoncer aux jalons existants ? La F-mesure possède des propriétés qui répondent souvent aux besoins mais il lui est impossible d'intégrer un niveau d'exigence sur le résultat espéré. Or dans certains secteurs, comme le domaine médical, les utilisateurs peuvent considérer qu'un résultat moyen est insuffisant ou inadapté. Il fallait donc trouver un indicateur capable de répondre à ce besoin, sans perdre les qualités des mesures existantes. Enfin, les mesures d'évaluation ne distinguent pas les tâches qu'elles mesurent : recherche d'information, classification ou catégorisation. Il fallait donc trouver également une méthodologie qui permette de nuancer en fonction de la finalité.

Pour cela, nous introduisons notre étude en tentant de préciser les principaux concepts utilisés : évaluation, classification, catégorisation, et recherche d'informations. Ces derniers termes feront l'objet d'une proposition de définition. La section 3 exposera l'état de l'art pour la problématique de l'évaluation, ainsi que les indicateurs existants. Enfin, après avoir analysé les propriétés de la F-Mesure (section 4), nous exposons une nouvelle approche de l'évaluation (section 5), adaptée aux objectifs et qui permet d'intégrer les exigences des juges.

2. Etymologie et définitions

Les termes 'classification' et 'catégorisation' ont des histoires et des origines très différentes. A ce jour, il semble persister une confusion entre catégorisation et classification, dans le langage courant voire philosophique. Aucune définition scientifique n'a pu être trouvée, hormis le Webster qui donne deux sens au mot classification : celui de 'taxonomie' (ou 'taxinomie') et celui de … catégorie.

Nous ne disposons donc que de l'historique de ces termes, leurs définitions successives par les philosophes et le dictionnaire de l'académie française. La première définition de la classification date de 1787 [Féraud 1787]. Le terme 'classification' apparaît pour la première fois dans la cinquième édition du dictionnaire de l'académie Française en 1798 [ACA 1798] sous la définition : « distribution en classes et suivant un certain ordre ». Le terme 'catégorisation' n'existe pas dans le dictionnaire de l'académie française, contrairement au mot 'catégorie'. On pourrait néanmoins le définir comme étant l'action de créer des catégories ou le résultat de cette action. Ce mot vient du grec katêgoria : « qualité attribuée à un objet ». Aristote définit les catégories comme étant les « espèces les plus générales de ce qui est signifié par un mot simple». Il regroupe dans un même ensemble des éléments proches et dénombre dix catégories. D'un point de vue philosophique, elles ne sont pas sans analogie avec les 'Principes', opposés deux à deux, dont certains Pythagoriciens voulaient que tout fût formé : fini et infini, pair et impair, repos et mouvement, etc.… D'autres mouvements philosophiques (les stoïciens, les logiciens de Port Royal, ou Kant par exemple) discuteront du nombre ou du bien fondé des catégories.

Compte tenu de l'historique de ces deux termes et de leur acception actuelle, **Nous proposons de définir la classification** comme étant l'action d'organiser un ensemble en structures ordonnées ou hiérarchisées, en classes préétablies ou non, ou le résultat de cette action **et la catégorisation** comme étant l'action (ou son résultat) de regrouper des éléments qui possèdent des caractéristiques communes. Cette définition se rapproche de la pensée aristotélicienne de par l'absence d'ordre et de hiérarchie. L'analogie avec les Pythagoriciens reste conservée car on maintient le mode de pensée booléen : une proposition d'affectation à une catégorie sera bonne ou 'non bonne' (valuation binaire 0/1). Par contre, dans une classification, il sera possible de quantifier ou valoriser l'écart entre la proposition et la valeur attendue. On pourra considérer une proposition comme étant partiellement bonne, et y associer une métrique.

Enfin, **la recherche d'informations** se différencie des classifications et catégorisations par le très grand nombre de réponses possibles, potentiellement infini, l'absence de référentiel, et donc l'obligation d'une évaluation humaine. L'application classique serait la réponse d'un moteur de recherche ou d'intelligence artificielle à une demande.

Comptes tenus de ces approches possibles, les méthodes d'évaluations et indicateurs peuvent être différents.

3. Etat de l'art

3.1 Qu'est que l'évaluation ?

L'évaluation consiste à mesurer la différence entre un résultat attendu et un résultat obtenu. Aucune métrique n'y est associée mais on utilise en général un indicateur compris entre 0 et 1 pour en faciliter l'interprétation.

Certains éléments sont très subjectifs et difficilement automatisables. Tefko Saracevic [SAR 70] met en avant l'importance du contexte sur la pertinence, ainsi que l'utilisateur (le juge). Par conséquent, si l'un de ces facteurs évolue, la notion de pertinence change également, et donc l'évaluation.

3.2 Les indicateurs et mesures : vers la F-mesure

Un système peut répondre à une requête selon le modèle suivant :

	Pertinent	Non pertinent	Total
Retrouvé (ou proposé)	a	b	a+b
Non retrouvé (ou non proposé)	c	d	c+d
	a+c	b+d	a+b+c+d=N

A partir de ce tableau de contingence, la communauté TALN calcule diverses mesures : la précision = $\frac{a}{a+b}$, le rappel = $\frac{a}{a+c}$, la pertinence= $\frac{a+d}{a+b+c+d}$, l'erreur = $\frac{b+c}{a+b+c+d}$, le taux de chute = $\frac{b}{b+d}$, le silence (=1-rappel) = $\frac{c}{a+c}$, la spécificité = $\frac{d}{b+d}$, le bruit = $\frac{b}{a+b}$, l'overlap = $\frac{a}{a+b+c}$ et la généralité = a/N

Au final, 4 mesures (a, b, c, d) génèrent 10 indicateurs de base, eux même combinés pour donner d'autres mesures. Dans la grande majorité des cas, on se limite à la précision et au

rappel. A partir de ces différentes mesures, plusieurs indicateurs de synthèse ont été créés, mais le plus célèbre est la F-Mesure de [Van Rijsbergen 79]:

Mesure F = ((1+ß²)*Précision*Rappel) / ((ß²*Précision)+Rappel), avec ß² = 1

Nous observons que cette mesure ne prend pas en compte la pertinence et fonctionne sur un modèle binaire : une réponse est bonne ou fausse. Dans le cas de l'évaluation d'un algorithme de classification, la F-mesure est fournie comme un résultat global. Elle est utilisée pour évaluer la performance d'une ou de chacune des classes [Dumais 98]. On distinguera la macro moyenne (*macro average*) de la micro moyenne (*micro average*). La macro moyenne calcule une moyenne globale, tandis que la micro moyenne calcule la moyenne des mesures de chaque catégorie [Yang 1999]. [Sebastiani 2004] propose la synthèse suivante :

	Micro averaging	Macro averaging
Precision	$\pi = \frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|} TP_i + FP_i}$	$\pi = \frac{\sum_{i=1}^{\|C\|} \pi_i}{\|C\|} = \frac{\sum_{i=1}^{\|C\|} \frac{TP_i}{TP_i + FP_i}}{\|C\|}$
Rappel	$\rho = \frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|} TP_i + FN_i}$	$\rho = \frac{\sum_{i=1}^{\|C\|} \rho_i}{\|C\|} = \frac{\sum_{i=1}^{\|C\|} \frac{TP_i}{TP_i + FN_i}}{\|C\|}$

3.3 Approche du Break Even Point

Comment calculer la précision globale (ou le rappel) d'un algorithme sur un ensemble de documents ? [Salton 1983] utilise une autre approche par la méthode des 11 points de moyenne (Break Even Point de [Joachims 98]), à partir de la meilleure valeur de précision et rappel dans chaque tranche de rappel variant de 0 à 1, par incrément de 0.10 et par extrapolation. [Moschitti 2003] expose en détail son mode de calcul. L'indicateur global consiste à trouver le couple où les deux mesures sont maximisées et leurs valeurs identiques.

3.4 Distance sémantique

[Rodriguez 2003] aborde la question de la distance sémantique de manière globale : le voisinage sémantique N d'une entité de classe est un jeu de classes se trouvant à une distance inférieure ou égale à un entier positif r, dénommé le radius du voisinage sémantique. [Roddick 2003] propose une formule générique de calcul de distance sémantique, mais qui nécessite toutefois de connaître la distance entre les nœuds. [JIN 2003] présente plusieurs mesures, en particulier une distance basée sur la corrélation entre classes. [Laurent 2005] utilise la distance de Levenshtein dans le projet industriel Lucene. [Budanitsky 2001] compare cinq distances sémantiques : la mesure de Hirst and St-Onge's [Hirst 1998] basée sur le nombre de changements de directions, la mesure de Leacock and Chodorow [Leacock 1998] basée sur le plus court chemin, la mesure de Resnik [Resnik 1995] fondée sur le niveau d'information commune, la mesure de Jiang et Conrath [Jiang 1997] calculée sur la probabilité d'apparition d'une classe, et enfin [Lin 1998] qui en est une autre déclinaison. Il obtient les meilleurs résultats avec la formule de Jiang and Conrath. [Jian-Yun Nie 2004] met l'accent sur le choix de la mesure de distance et cite le cas d'une désambiguïsation détériorée par l'usage d'une distance calculée selon le nombre de liens entre les synsets dans l'arbre de hypernyme-hyponyme. Il met en cause l'utilisation du nombre de liens comme mesure de distance sémantique. Dans le domaine des classifications médicales, [Bousquet 2003] utilise une distance sémantique, en projetant les diagnostics sur les axes de la SNOMED.

3.5 Approches humaines ou non binaires

[Voohrees 2003] utilise un score spécifique pour analyser les résultats de la conférence TREC : $\frac{1}{Q}\sum_{i=1}^{Q}\frac{n}{i}$, ou n représente le nombre de réponses exactes au rang i, et Q le nombre de questions. Ce score privilégie donc le fait de proposer en premier les bonnes réponses. Elle souligne la mauvaise reproductibilité de l'évaluation entre experts humains : à la question « Who is Tom Cruise married to? », certains jugent acceptèrent « Kidman » comme bonne réponse, mais d'autres la considéraient comme inexacte. Quelques jugent ont validé « actress Nicole Kidman », d'autres non. Elle demande : quand une bonne réponse est « 5 octobre 1996 », pouvons nous considérer comme faux un système qui donnerait « 04 octobre 1996 » ? [Dimitrova 2004] a analysé les classifieurs de documents web en se basant sur l'évaluation humaine, et mis en évidence l'émergence de la dimension cognitive. Elle illustre parfaitement l'importance du juge et la complexité de cette approche.

Les documents sont, selon l'inventeur du web, "destinés aux humains plutôt que des données et informations qui peuvent être analysées automatiquement" [Berners-Lee et al. 2001], le défi est justement d'extraire automatiquement de l'information de ces documents écrits en langage naturel. Berners y répond en introduisant le semantic web, pour faciliter l'exploitation du web par les outils.

4. Analyse de la F-Mesure

Nous avons démontré dans un premier temps que la F-mesure correspond en fait à une moyenne harmonique de la précision et du rappel, puis observé ses propriétés. Le paramètre ß permet de pondérer la précision ou au rappel, mais vaut généralement 1. La mesure devient : (2*Précision*Rappel) / (Précision+Rappel)

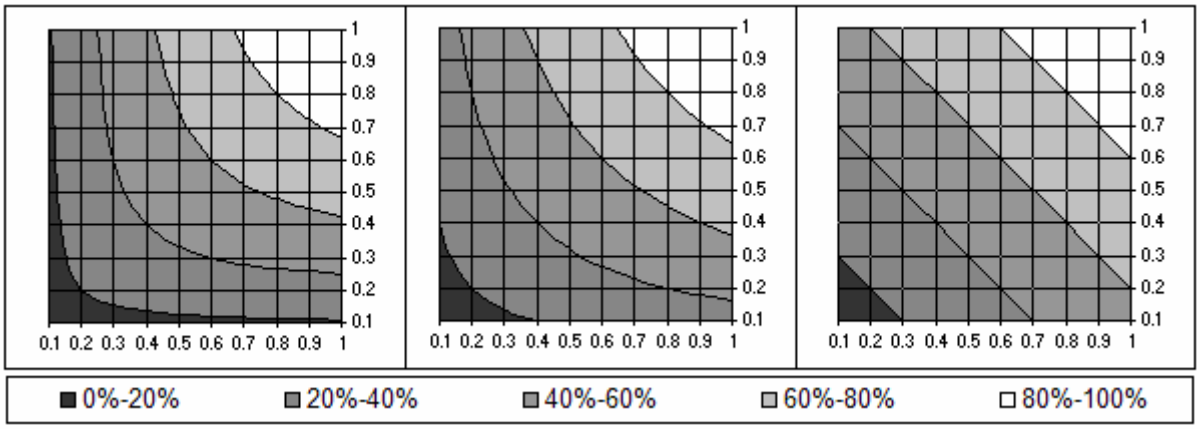
L'avantage de ce choix est que lorsque la précision est égale au rappel, on obtient : Précision = Rappel = F1-mesure. Ceci facilite la lecture et on recherche à maximiser la F-mesure en maximisant simultanément la précision et le rappel (dans l'esprit du calcul du 'Break Even Point'). En effet, il deviendrait difficile d'apprécier un algorithme qui fournirait une bonne précision et un mauvais rappel ou inversement.

Considérons la moyenne harmonique M de la précision (P) et du rappel (R) : $\frac{2}{M}=\frac{1}{P}+\frac{1}{R}$ donc $\frac{2}{M}=\frac{P+R}{P*R}$, soit: $\frac{M}{2}=\frac{P*R}{P+R}$ et finalement : $M=\frac{2*(P*R)}{P+R}=F1$

Nous constatons que la F1-mesure n'est que la moyenne harmonique de la précision et du rappel. Rien ne justifie cela. Néanmoins, la moyenne harmonique possède une qualité qui est de baisser fortement lorsque l'un de ses paramètres est bas et d'être majorée lorsque les deux paramètres sont proches entre eux tout en étant élevés. Cette propriété pénalise donc les algorithmes qui dégraderaient la précision ou le rappel, ou l'un exclusivement au détriment de l'autre. Prenons la F1-mesure et démontrons cette propriété : nous avons F1=2*P*R/(P+R), avec précision=P et rappel=R. Posons S=P+R et D=P-R. Notre problème devient donc : comment impacter la F-mesure pour qu'elle soit optimale quand on maximise S (donc la précision ET le rappel) en minimisant D (en les conservant proches). Nous avons : $S^2-D^2=(S+D)(S-D)=(P+R+P-R)\times((P+R)-(P-R))=2P\times 2R=4PR$

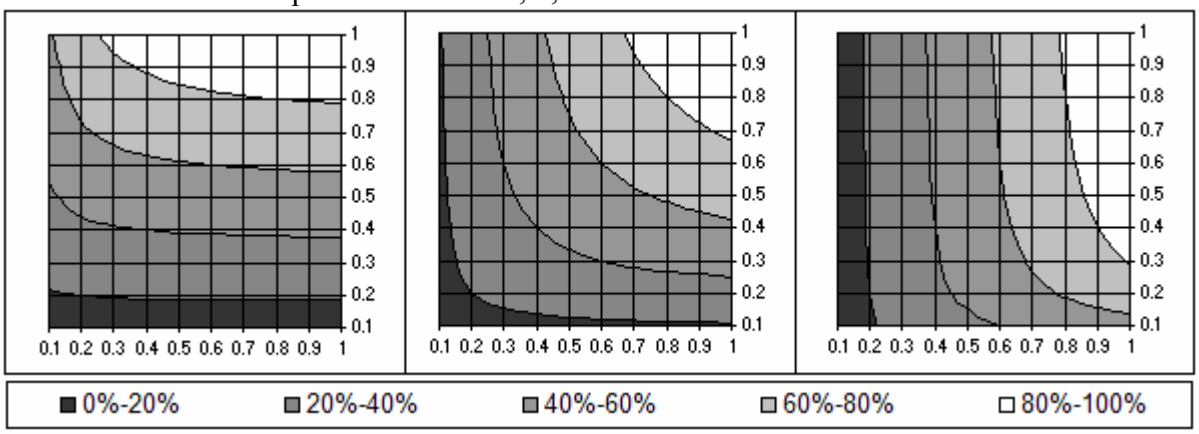
En conclusion, on a : $F1=\frac{2*P*R}{P+R}=\frac{S^2-D^2}{2S}=\frac{S}{2}-\frac{D^2}{2S}$; ceci explique pourquoi la Fmesure

croît quand S augmente, et décroît quand D augmente. Nous pouvons observer cette propriété sur les graphiques suivants, où l'on compare la moyenne harmonique avec la moyenne géométrique et arithmétique :



Nous observons que la surface noire (mesure d'évaluation comprise entre 0 et 0.20, donc mauvais score) est plus étirée vers les extrêmes, ce qui traduit bien le fait que si l'un des composants est faible, la mesure d'évaluation est faible.

La F_n-mesure possède une autre propriété : elle permet d'accorder plus de poids à la précision ou au rappel. Pour cela, on modifie le dénominateur original (P+R) en ((ß²*Précision)+Rappel). Le résultat de ce calcul est compris entre 0 et ß²+1. Pour que rester dans l'intervalle [0,1], il faut donc multiplier le numérateur par (ß²+1). Cela nous donne la formule de la Fmesure. Nous pouvons observer la pertinence de cette propriété, en donnant 3 valeurs successives au paramètre ß² : 0.3, 1, et 3.



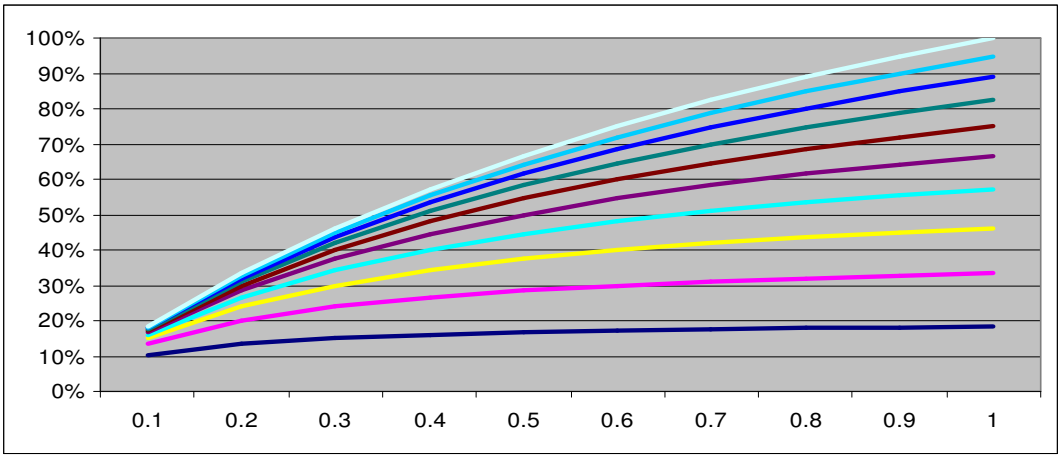
5. Propositions de nouveaux indicateurs : vers la K-mesure

Dans la section 1, nous avons tenté de définir la classification et de la dissocier de la catégorisation et de la recherche d'informations. Nous allons maintenant étudier quels pourraient être les indicateurs les plus riches en fonction de la nature des tâches évaluées.

Cas des catégorisations

Dans le cas d'une catégorisation, il n'y a aucune distance entre une proposition et la solution désirée. Nous n'avons d'autre possibilité que celle d'utiliser un système binaire, soit 0 ou 1 selon que le résultat est faux ou vrai. La F-mesure possède des propriétés très intéressantes exposées dans la section précédente, mais [Voorhees 2003], [Dimitrova 2004] ont mis en évidence l'importance du juge. [Van Rijsbergen 79] y répond très partiellement en

introduisant le paramètre β, mais aucune variable ne permet à l'utilisateur de mettre un niveau d'exigence en terme de précision ou de rappel. Sur le graphique suivant (chaque courbe représente un rappel variant de 0.1 à 1 et affiche la Fmesure résultante), nous constatons que le faible niveau de précision (ou rappel) abaisse la courbe, mais qu'au-delà de 0.4, le système devient pratiquement linéaire.

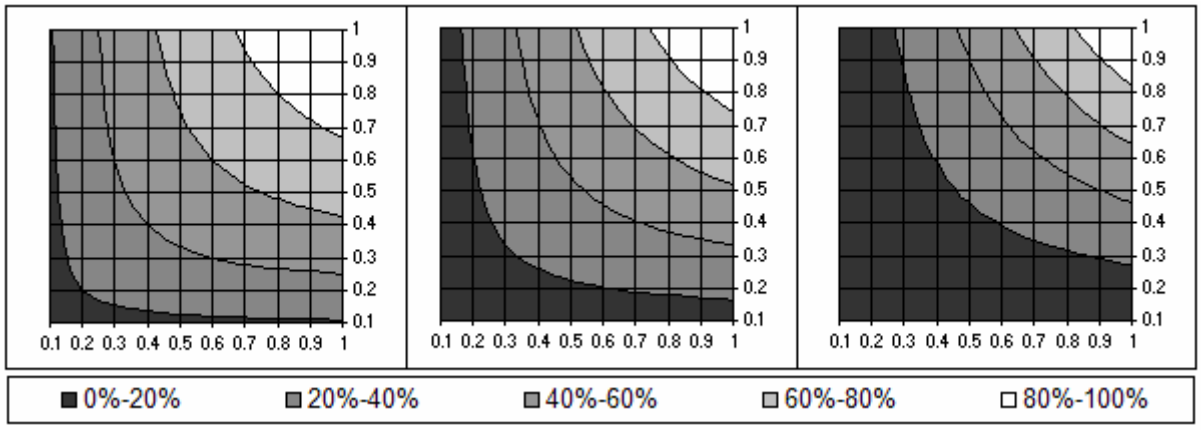


Or un utilisateur (le juge de Saracevic) peut considérer une précision ou un rappel de 0.5 comme méritant une mauvaise évaluation. La Fmesure ne permet pas de traduire ce besoin.

De manière empirique, nous avons trouvé une formule qui permet d'intégrer ces contraintes dans la F-mesure, et introduisons la **K-Mesure** :

$$\textbf{K-Mesure} = (1+ß^2)*(\text{Précision}*\text{Rappel})^{\alpha} / ((ß^2*\text{Précision})+\text{Rappel})$$

Etudions maintenant les propriétés de la K-Mesure. En premier lieu, nous constatons que si α=1, nous obtenons la F-mesure, dans sa version la plus complète. En posant α=1 et ß²=1, nous obtenons la F-mesure dans son acception la plus usuelle. La K-mesure est donc un sur ensemble de la F-mesure. Ainsi, tout l'historique peut être conservé, ce qui facilite les comparaisons. Observons maintenant les propriétés de la K-mesure, en faisant varier le paramètre α (1.0, 1.2, 1.6), avec ß²=1.



Nous constatons que plus le paramètre α augmente, plus le niveau d'exigence du juge grandit. Par exemple, avec P=R=0.4, la Fmesure vaut 0.4, alors que la K-mesure vaut trois fois moins (0.13) pour α =1,6. Ce résultat sera donc considéré comme mauvais alors que la F-mesure le considère comme moyen. Nous pouvons donc ainsi formaliser un niveau d'exigence du juge.

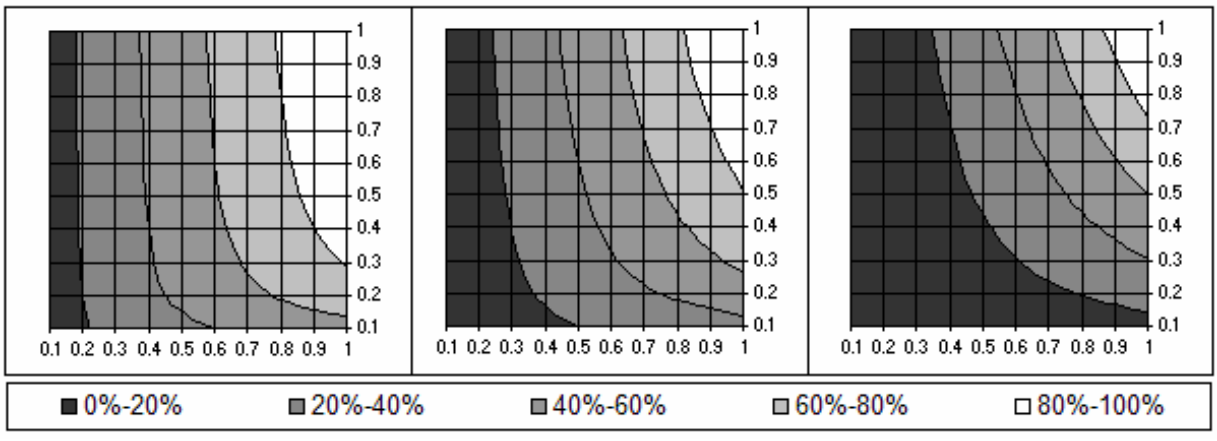
Voici les valeurs obtenues avec le paramètre α (1.0, 1.2, 1.6), et ß²=1. Nous faisons varier la précision et le rappel de 0 à 1 par pas de 0.1, ce qui nous donne les valeurs suivantes pour la K-mesure :

α =	**1**								ß²=	**1**
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	10%	13%	15%	16%	17%	17%	18%	18%	18%	18%
0.2	13%	20%	24%	27%	29%	30%	31%	32%	33%	33%
	15%	24%	30%	34%	38%	40%	42%	44%	45%	46%
0.4	16%	27%	34%	40%	44%	48%	51%	53%	55%	57%
0.5	17%	29%	38%	44%	50%	55%	58%	62%	64%	67%
0.6	17%	30%	40%	48%	55%	60%	65%	69%	72%	75%
0.7	18%	31%	42%	51%	58%	65%	70%	75%	79%	82%
0.8	18%	32%	44%	53%	62%	69%	75%	80%	85%	89%
0.9	18%	33%	45%	55%	64%	72%	79%	85%	90%	95%
1	18%	33%	46%	57%	67%	75%	82%	89%	95%	100%

α =	**1.2**								ß²=	**1**
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	4%	6%	7%	8%	9%	10%	10%	11%	11%	11%
0.2	6%	11%	14%	16%	18%	20%	21%	22%	23%	24%
0.3	7%	14%	19%	22%	26%	28%	31%	33%	35%	36%
0.4	8%	16%	22%	28%	32%	36%	39%	42%	45%	48%
0.5	9%	18%	26%	32%	38%	43%	47%	51%	55%	58%
0.6	10%	20%	28%	36%	43%	49%	54%	59%	64%	68%
0.7	10%	21%	31%	39%	47%	54%	61%	66%	72%	77%
0.8	11%	22%	33%	42%	51%	59%	66%	73%	79%	85%
0.9	11%	23%	35%	45%	55%	64%	72%	79%	86%	93%
1	11%	24%	36%	48%	58%	68%	77%	85%	93%	100%

α =	**1.6**								ß²=	**1**
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	1%	1%	2%	2%	3%	3%	4%	4%	4%	5%
0.2	1%	3%	4%	6%	7%	8%	10%	11%	12%	13%
0.3	2%	4%	7%	10%	12%	14%	16%	19%	21%	22%
0.4	2%	6%	10%	13%	17%	20%	24%	27%	30%	33%
0.5	3%	7%	12%	17%	22%	26%	31%	36%	40%	44%
0.6	3%	8%	14%	20%	26%	33%	38%	44%	50%	55%
0.7	4%	10%	16%	24%	31%	38%	46%	53%	60%	66%
0.8	4%	11%	19%	27%	36%	44%	53%	61%	70%	78%
0.9	4%	12%	21%	30%	40%	50%	60%	70%	79%	89%
1	5%	13%	22%	33%	44%	55%	66%	78%	89%	100%

Observons à présent si la possibilité d'avantager la précision au détriment du rappel (ou inversement) est conservée, toujours avec α =1.0, 1.2, 1.6, mais ß²=3 :



Nous constatons qu'effectivement, cette possibilité est conservée. Nous obtenons la même propriété en faisant varier l'autre axe (soit des valeurs du paramètre β inférieures à 1).
Nous pouvons observer le détail des valeurs :

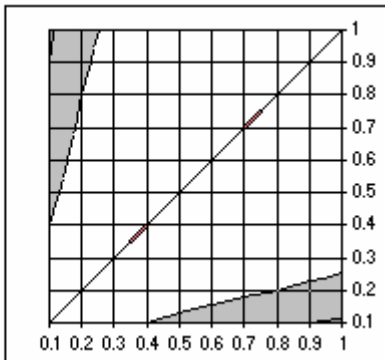
α =	1								Beta=	3
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	10%	18%	25%	31%	36%	40%	44%	47%	50%	53%
0.2	11%	20%	29%	36%	43%	50%	56%	62%	67%	71%
0.3	11%	21%	30%	39%	47%	55%	62%	69%	75%	81%
0.4	11%	21%	31%	40%	49%	57%	65%	73%	80%	87%
0.5	11%	21%	31%	41%	50%	59%	67%	75%	83%	91%
0.6	11%	21%	32%	41%	51%	60%	69%	77%	86%	94%
0.7	11%	22%	32%	42%	51%	61%	70%	79%	88%	96%
0.8	11%	22%	32%	42%	52%	62%	71%	80%	89%	98%
0.9	11%	22%	32%	42%	52%	62%	72%	81%	90%	99%
1	11%	22%	32%	43%	53%	63%	72%	82%	91%	100%

a=	1.2								Beta=	3
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	4%	8%	12%	16%	20%	23%	26%	28%	31%	33%
0.2	5%	11%	16%	22%	27%	33%	38%	43%	47%	52%
0.3	5%	12%	19%	25%	32%	39%	45%	52%	58%	64%
0.4	6%	13%	20%	28%	35%	43%	50%	58%	65%	72%
0.5	6%	13%	21%	30%	38%	46%	55%	63%	71%	79%
0.6	6%	14%	22%	31%	40%	49%	58%	67%	76%	85%
0.7	6%	15%	23%	32%	42%	51%	61%	70%	80%	89%
0.8	7%	15%	24%	34%	43%	53%	63%	73%	83%	93%
0.9	7%	15%	25%	35%	45%	55%	65%	76%	86%	97%
1	7%	16%	25%	35%	46%	56%	67%	78%	89%	100%

a=	1.6								Beta=	3
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	1%	2%	3%	4%	6%	7%	9%	10%	12%	13%
0.2	1%	3%	5%	8%	11%	14%	17%	20%	24%	27%
0.3	1%	4%	7%	11%	15%	19%	24%	29%	34%	39%
0.4	2%	5%	9%	13%	19%	24%	30%	37%	43%	50%
0.5	2%	5%	10%	16%	22%	29%	36%	44%	52%	60%
0.6	2%	6%	11%	18%	25%	33%	41%	50%	59%	69%
0.7	2%	7%	12%	19%	27%	36%	46%	56%	66%	77%
0.8	2%	7%	14%	21%	30%	40%	50%	61%	73%	85%
0.9	3%	8%	15%	23%	32%	43%	54%	66%	79%	93%
1	3%	8%	16%	25%	35%	46%	58%	71%	85%	100%

Enfin, nous pouvons donner au paramètre α une valeur comprise en 0.5 et 1, à condition que l'on ait β =1 (sinon, le résultat pourrait devenir supérieur à 1). La valeur ne peut en aucun cas être inférieure à 0.5 car les propriétés désirées ne seraient plus respectées. Une valeur remarquable est pourtant intéressante : α = 0.5 :

α =	0.5								Beta=	1
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	100%	94%	87%	80%	75%	70%	66%	63%	60%	57%
0.2	94%	100%	98%	94%	90%	87%	83%	80%	77%	75%
0.3	87%	98%	100%	99%	97%	94%	92%	89%	87%	84%
0.4	80%	94%	99%	100%	99%	98%	96%	94%	92%	90%
0.5	75%	90%	97%	99%	100%	100%	99%	97%	96%	94%
0.6	70%	87%	94%	98%	100%	100%	100%	99%	98%	97%
0.7	66%	83%	92%	96%	99%	100%	100%	100%	99%	98%
0.8	63%	80%	89%	94%	97%	99%	100%	100%	100%	99%
0.9	60%	77%	87%	92%	96%	98%	99%	100%	100%	100%
1	57%	75%	84%	90%	94%	97%	98%	99%	100%	100%



Nous observons que cette valeur permet de privilégier fortement l'indicateur lorsque l'on a précision P = rappel R, ce qui correspond à l'approche du 'Break Even Point'.

En conclusion, la K-mesure possède des propriétés qui la rendent intéressante :

- Elle est un sur ensemble de la Fmesure et en conserve les qualités
- Elle permet de représenter un niveau d'exigence du juge
- Elle permet de représenter également le Break Even Point

Il s'agit donc d'une formule de convergence, et d'un sur ensemble des différentes approches utilisées à ce jour.

Cas des classifications

Comme proposé en section 1, une classification se distingue d'une catégorisation par le fait qu'il est possible de mettre en œuvre une mesure de distance entre les classes. Dans le cas d'une classification hiérarchique, la mesure de la précision et du rappel présente un paradoxe : plus la hiérarchie est fine, plus la mesure de l'évaluation sera mauvaise. Prenons le cas de la CIM 10 (classification internationale des maladies) : au niveau 4, nous disposons de 15000 codes possibles, ce qui permet d'obtenir un niveau d'individualisation déjà très fin. Mais si l'on fait une estimation sur le niveau 5, il faut alors trouver le bon code parmi 51 000 possibilités. La probabilité de trouver le bon code diagnostic devient alors plus faible et précision et rappel seront également plus faibles.

En généralisant ce principe, on peut en déduire que la précision et le rappel convergent vers 0 lorsque le niveau de hiérarchie croît. Mais on peut alors considérer qu'une réponse donnée à un niveau très bas de hiérarchie avec juste une erreur sur le dernier niveau de hiérarchie est pratiquement correcte. Dans le cas d'une Fmesure, elle serait considérée comme fausse.

[Budanitsky 2001] a démontré que les meilleurs résultats étaient obtenus avec la mesure de Jiang et Conrath :

$$\text{Soit } d=Dist_{jc}(c1 :c2) = 2 \log(p(lso(c1 :c2)))-(\log(p(c1))+\log(p(c2)))$$

Avec lso(c1 :c2)= plus grand sous groupe commun.

Si l'on nomme 'd' cette distance (avec d=1 si les classes sont distantes), alors la précision et le rappel peuvent alors se définir comme suit :
Précision = a / B et Rappel = a / c
a = Nombre de classes pertinentes et proposées (donc correctement classifiées)
B= Elément proposé mais non pertinent, on prend la mesure 'd' de distance avec la solution la plus proche. Puis 1-d pour amener B vers 1 quand la distance est faible.
c = Nombre d'éléments non proposés mais pertinents

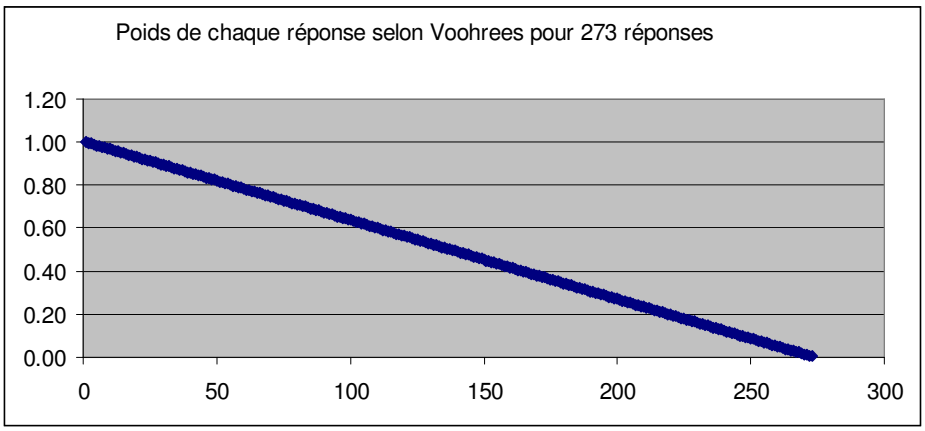
Il devient alors possible d'utiliser la K-mesure.

Cette proposition d'utiliser la hiérarchie pour affiner la qualité de l'évaluation augmentera mécaniquement les valeurs de précision et de rappel mais sera plus proche de la réalité. Or c'est bien l'objectif recherché par la mise en œuvre d'un indicateur d'évaluation. En utilisant cette méthodologie et la K-mesure, on peut parfaitement exprimer à la fois un niveau d'exigence et la notion de proximité.

Cas des recherches d'informations (IR)

La recherche d'informations se différencie des classifications et catégorisations par le très grand nombre de réponses possibles. L'application classique serait celle d'un moteur de recherche.

Pour établir un indicateur, nous sommes partis de la formule de score donnée par [Voohrees 2003] $\frac{1}{Q}\sum_{i=1}^{Q}\frac{n}{i}$, ou n représente le nombre de réponses exactes au rang i, et Q le nombre de questions. Ce qui donne la représentation linéaire suivante :

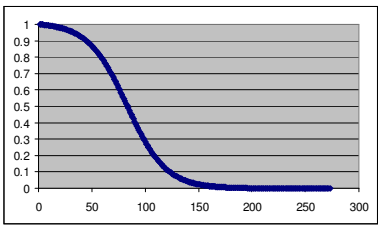


Pour pouvoir représenter un niveau d'exigence du juge (exemple : "je veux absolument que les réponses les plus pertinentes soient sur la première page de réponse, soit 30 réponses"), il faut moduler cette approche booléenne et linéaire par une fonction sigmoïdale. Après des recherches empiriques, nous avons trouvé un coefficient de pondération Wi qui répond à ce besoin :

$$w_i = \frac{1+e^{(-k*(1-l))}}{1+e^{(-k*(((N-i+1)/N)-l))}} = \frac{1+e^{(l-k)}}{1+e^{\left(-k\times\left(\frac{N-i+1}{N}\right)-l\right)}}$$

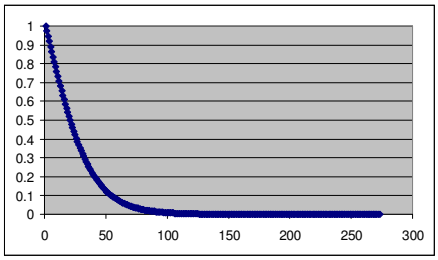
avec k et l, deux paramètres (par défaut k=15, l=0.7), N représente l'effectif total, i le rang de la réponse évaluée.

Observons les propriétés de cette équation en faisant varier les deux paramètres k et l (pour cette illustration, nous avons pris N=273)
k=15, l=0.7

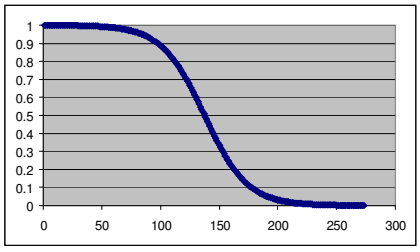


Nous observons que si la réponse attendue n'apparaît pas dans les 50 premières réponses proposées, le score est fortement minoré, voire quasi nul au delà de 150. Le paramètre l déplace le point d'inflexion, alors qu le paramètre k modifie le niveau d'inclinaison :

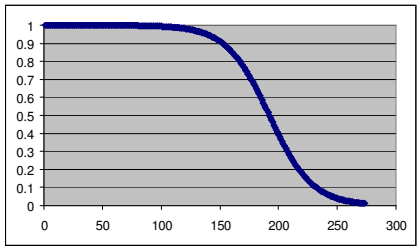
k=15, l=1



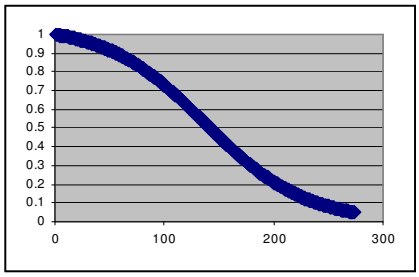
k=15, l=0.5



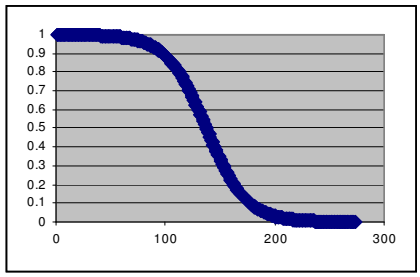
k=15, l=0.2



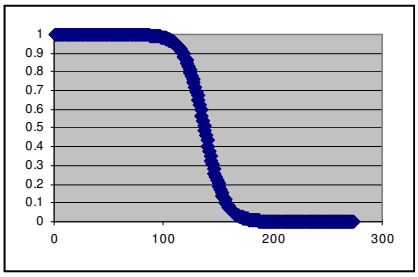
k=6, l=0.5



k=15, l=0.5



k=30, l=0.5



Les deux paramètres permettent donc de générer n'importe quel niveau d'exigence. Ce score privilégie donc le fait de proposer en premier les bonnes réponses.

Pour calculer l'indicateur final, il suffit de multiplier le poids de la réponse Wi par son niveau de pertinence. Dans le cas d'un traitement automatique, nous pouvons utiliser une logique booléenne : coder 1 pour une bonne réponse, sinon 0. Dans le cas d'une évaluation humaine, le juge peut donner un pourcentage

L'indice global d'évaluation devient alors : $$Evaluation = \frac{\sum_{i=1}^{N} \frac{Pertinence_i \times 1 + e^{(l-k)}}{1 + e^{(-k \times \left(\frac{N-i+1}{N}\right) - l))}}}{\sum_{i=1}^{N} w_i}$$

6. Conclusion

Dans cet article, nous avons en premier lieu tenté de définir classification et catégorisation. Ceci a mis en évidence que dans le premier cas il était possible de mesurer une distance entre les classes, alors que dans le second nous sommes dans un système binaire.

L'étude de la F-mesure nous a permis de montrer que cette mesure, créée il y a 25 ans, s'est imposée comme standard d'évaluation. Depuis, les besoins ont évolué, sans remise en cause des indicateurs. L'analyse de la problématique nous a permis d'introduire une nouvelle mesure : la K-mesure, qui est un sur ensemble de la F-mesure, capable d'intégrer des niveaux d'exigence de l'utilisateur. Nous avons également examiné comment il était possible d'utiliser la K-mesure pour les classifications, tout en intégrant la notion de distance entre le résultat attendu et le résultat obtenu. Enfin, nous proposons une mesure spécifique aux problèmes de recherche d'information qui privilégie le fait de donner en premier les bonnes réponses, avec également la possibilité de moduler le niveau d'exigence.

La K-mesure présente les avantages suivants : d'une part, elle représente une méta mesure de convergence entre la F-mesure de Van Rijsbergen et le break even point de Joachims. D'autre part, elle possède des propriétés mathématiques qui lui permettent de faire un indicateur de synthèse à partir de n'importe quelles autres mesures, dès lors que l'on souhaite les maximiser ensembles. Enfin, elle permet d'intégrer l'approche de Saracevic et de formaliser des niveaux d'exigence (les juges). Il s'agit donc d'une mesure qui est à la convergence de ses trois approches sans en dénaturer le sens.

En conclusion, cette étude permet de mettre en place deux indicateurs d'évaluation basés sur les indicateurs actuels, mais permettant d'intégrer des niveaux d'exigence des juges.

Dans nos futurs travaux, nous envisageons d'expérimenter ces mesures, en particulier d'étudier comment elles peuvent impacter les autres mesures.

7. Remerciements

Le travail présenté dans cet article est partiellement financé par le MENRT dans le cadre du projet RNTS Rhéa.

Nous tenons à remercier tout particulièrement :

- l'association OutcomeRea et son président Monsieur le Professeur Jean François Timsit pour leur aide dans le projet CIREA
- Monsieur le Professeur Jacky AKOKA et le laboratoire Cedric du CNAM (Conservatoire National des Arts et Métiers)
- Messieurs Pierre KEBAILI et Jacques TONNER (CRAMIF : Caisse Régionale d'Assurance Maladie d'Ile de France) pour leur soutien.

8. Bibliographie

[Aca 1798] Dictionnaire de L'Académie française, 5th Edition (1798), p767 http://colet.uchicago.edu

[Berners-Lee et al. 2001] T. Berners-Lee, J. Hendler, O. Lassila : "The Semantic Web", Scientific American. 2001

[Bousquet 2003] C Bousquet, MC Jaulent, C Le Bozec, P Degoulet. : "New terminology services based on term comparison using semantic definitions and similarity computation". MIE 2003

[Budanitsky 2001] A. Budanitsky and G. Hirst : "Semantic distance inWordNet: An experimental, application-oriented evaluation of five measures" Department of Computer Science Univ. of Toronto

[Dimitrova 2004] Maya Dimitrova, N. Kushmerick, I. Terziev and A. Gegov : "Web Users and Web Document Classifiers Emergent Cognitive Phenomena"; Proceedings of ISKO 2004 The 8th International Conference of the International Society for Knowledge Organization. http://www.cs.ucd.ie/staff/nick/home/research/download/dimitrova-isko2004.pdf

[Dumais et al. 1998] S. Dumais, J. Platt, D. Heckerman, M. Sahami : "Inductive Learning Algorithms and Representations for Text Categorization". Proceedings of the seventh International Conference on Information and Knowledge Management (CIKM' 98), 148-155, 1998.

[Féraud 1787] J-F. Féraud: « Dictionnaire critique de la langue française » (Marseille, Mossy 1787-1788). CLASSIFICATION (Page A458a). http://colet.uchicago.edu/

[Hirst 1998] G. Hirst and D. St-Onge. 1998 : "Lexical chains as representations of context for the detection and correction of malapropisms". In Fellbaum 1998, pp. 305–332.

[Jiang 1997] Jay J. Jiang, David W. Conrath. 1997. "Semantic similarity based on corpus statistics and lexical taxonomy". Proceedings of Internat. Conf. on Research in Computational Linguistics, Taiwan.

[Jian-Yun Nie 2004] http://www.iro.umontreal.ca/~nie/IFT6255/NLP.html

[JIN 2003] X. Jin and B. Mobasher, "Using Semantic Similarity to Enhance Item-Based Collaborative Filtering", in Proceedings of The 2nd IASTED International Conference on Information and Knowledge Sharing, Scottsdale, Arizona, November 2003.

[Joachims 1998] T. Joachims. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". Proceedings of the Tenth European Conference on Machine Learning (ECML'98).

[Laurent 2005] Jean-Marc LAURENT : "projet Lucene". http://connaissances.neuf.fr/rc.htm et http://www.merriampark.com/ld.htm

[Leacock 1998] Claudia Leacock and Martin Chodorow. "Combining local context and WordNet similarity for word sense identification". In Fellbaum 1998, pp. 265–283.

[Lin 1998] Dekang Lin. 1998. "An information-theoretic definition of similarity". In Proceedings of the 15th International Conference on Machine Learning,Madison, WI.

[Moschitti 2003] A. Moschitti "A study on optimal parameter tuning for Rocchio Text Classifier". University of Rome, 25th BCS-IRSG European Conference on Information Retrieval Research (ECIR'03). www.utdallas.edu/~axm020400/ECIR03.pdf

[Resnik 1995] Philip Resnik. “Using information content to evaluate semantic similarity”. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, p 448–453, Montreal.

[Roddick 2003] J. F. Roddick, K. Hornsby, D. de Vries : "A Unifying Semantic Distance Model for Determining the Similarity of Attribute Values", Proceedings of the Twenty-Sixth Australasian Computer Science Conference (ACSC2003), Adelaide, Australia, 2003

[Rodriguez 2003] A. Rodriguez, M. J. Egenhofer : "Determining semantic similarity among entity classes from different ontologies". IEEE Transactions on Knowledge and Data Engineering, 2003.

[Salton 1983] G. Salton, M.J. Mc Gill : “introduction to modern information retrieval”. Mc Graw Hill Computer Science series. Mc Graw Hill, New York, 1983.

[SAR 70] Tefko Saracevic, “Introduction to Information Science”, 111-151. New York: R.R. Bowker, 1970. Chap. 3 : The concept of "relevance" in information science: A historical review.

[Sebastiani 2004] F. Sebastiani : “Text Categorization” - In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK. Forthcoming.

[Sparck Jones. 2001] K. Sparck Jones. “Automatic language and information processing: Rethinking evaluation”. Natural Language Engineering, 7(1):29–46. 2001

[Van Rijsbergen 79] K. Van Rijsbergen, “Information Retrieval”, (2nd Ed.) Butterworths, London. www.dcs.gla.ac.uk/Keith/Preface.html

[Voorhees 2003] E. M. Voorhees : “Evaluating the Evaluation: Edmonton”, May-June 2003. Main Papers , pp. 181-188. Proceedings of HLT-NAACL 2003

[Yang 1999] Y. Yang : “An evaluation of statistical approaches to text categorization”. Journal of Information Retrieval, Vol 1, No. 1/2, pp 67--88, 1999.