Calcul de distance inter-documents par approche mots-clés

Alain Simac-Lejeune

▶ **To cite this version:**

Alain Simac-Lejeune. Calcul de distance inter-documents par approche mots-clés. 13e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances, Jan 2013, Toulouse, France. pp.501-506. hal-01059048

HAL Id: hal-01059048
https://hal.science/hal-01059048

Submitted on 29 Aug 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Calcul de distance inter-documents par approche mots-clés

Alain Simac-Lejeune*

*Compilatio
alain@compilatio.net,
http://www.compilatio.net

1 Extraction des mots-clés

Le terme *mot-clé* désigne de manière générale un mot qui a une importance particulière qu'on définit plus particulièrement en informatique comme un descripteur qui permet de caractériser l'information contenue dans un document. La principale application des mots-clés est la facilitation à la recherche de documents. La méthode la plus 'traditionnelle' pour extraire des mots-clés est la méthode fréquentielle (Lee et Baik, 2004) qui consiste pour chaque mot du texte à calculer sa fréquence d'apparition dans le texte.

2 Le calcul de distance

On définit l'écart maximal de distribution E qui représente le plus fort Δ observé dans la distribution courante entre les fréquences f (la fréquence étant le rapport entre le nombre d'occurrences d'un mot sur le nombre de mots total dans le document) des mots-clés du document i et celles des mots-clés du document j comme $E = max(||max(f_i) - min(f_j)||, ||max(f_j) - min(f_i)||)$. On définit également α comme le nombre de mots clés communs et $\overline{\alpha}$ le nombre de mots-clés différents entre deux documents et n le nombre de mots-clés utilisés. La distance entre deux documents proposé est composée d'un critère de convergence $C \in [0, E]$ qui vaut Δf, d'un critère de divergence $\overline{\alpha} * E$ normalisé par la taille de la distribution choisie E. La distance est ainsi définie sur l'intervalle $[0, 1]$:

$$\mathscr{D} = \frac{\sum^{\alpha} \Delta o + \overline{\alpha} * E}{n * E} \tag{1}$$

3 Utilisation et évaluation de la métrique

La base de tests est composée de 100 textes d'origine variée (publications, mémoires...) de 2500 mots environ présentant parfois des plagiats. On cherche pour chaque texte, les dix premiers mots-clés sur l'ensemble du texte puis sur des sous-parties du document. On évalue la qualité en terme de précision et de rappel. Une étude préliminaire a permet d'évaluer l'influence du seuil de distance utilisé dans la recherche de documents similaires en utilisant

	100		200		500		1000		2500	
2 mots-clés	0.63	1	0.60	1	0.61	1	0.71	1	0.74	1
3 mots-clés	0.93	1	0.94	1	0.95	1	0.97	0.98	0.96	0.97
5 mots-clés	0.97	0.92	0.99	0.93	0.98	0.92	0.99	0.91	1	0.90
Mots	0.92	1	0.89	0.99	0.95	0.99	0.91	0.98	0.89	0.94
Chaine	0.95	1	0.93	1	0.96	1	0.94	0.95	0.95	0.96
TF-IDF	0.97	1	0.96	1	0.96	1	0.95	0.96	0.96	0.96
Nous	0.92	1	0.94	1	0.96	1	0.97	0.98	0.96	0.97

TAB. 1 – *Précision/Rappel en utilisant la recherche de similitudes proposée, par comparaison de mots, de chaînes, par TF-IDF en utilisant de 1 à 10 mots-clés.*

3 mots-clés : l'utilisation d'une distance de 0.08 permet de conserver une bonne précision et d'obtenir une valeur intéressante de rappel.

Les résultats permettent d'évaluer le positionnement de notre méthode par rapport à la recherche de similitudes par comparaison de mots et par comparaison de chaînes qui est globalement **meilleure en precision/rappel** que l'approche par mots et très proche bien qu'en moyenne un peu meilleure que l'approche par chaînes.

4 Conclusion

La distance inter-documents proposée montrent des résultats proches voir meilleurs que les approches classiques de comparaison. Cette métrique est rapide à calculer et diminue le nombre de données à utiliser pour la comparaison de documents mais nécessite l'utilisation de textes d'au moins 100 mots pour donner des résultats utilisables voir 500 mots pour devenir satisfaisantes.

Références

Lee, J.-W. et D.-K. Baik (2004). A model for extracting keywords of document using term frequency and distribution. In *Computational Linguistics and Intelligent Text Processing*, Volume 2945, pp. 437–440.

Summary

The notion of keywords is a principal element in the text analysis because they can summarize information from a text from several thousand to a few words. While searching for documents by keywords is a process that works quite well, the comparison of texts is a task that does not use these keywords, which is based only on the full text. In this paper, we define a metric to calculate the distance between two documents or parts of documents based on the extracted keywords, and we validated on a heterogeneous databases. The objective is to provide a robust and effective for document comparison based solely on a few keywords.