A Comparison of Methods for Automatic Document Classification

The process of assigning keyphrases to a document based on its contents is known as automatic keyphrase extraction (AKE). Previous research, such as (Sood et al, 2007), has shown that in significant numbers of instances, keyphrases supplied by authors are not appropriate for the document with which they are associated. Often they include keyphrases that are classificatory rather than explanatory (e.g., "University of Poppleton" instead of "Knowledge Discovery in Databases") or were not updated when the document's focus changed after its conception.

Numerous distinct systems have been presented which attempt this task, but mostly all use different corpora to evaluate their results. This makes comparing the effectiveness and efficiency of each system difficult, and therefore impairing the ease of incorporating useful elements of algorithms in future work.

This paper describes work with an aim was to implement a small selection of different methods for AKE, including the C-Value (Frantziy et al, 2000), an *n*-gram method (Hussey et al, 2011), and basic statistical methods (tf and tf*idf). These methods were then evaluated against a range of corpora to compare on issues such as performance, quality of results, and setting a benchmark against which to compare new systems.

References:

S.C. Sood, S.H. Owsley, K.J. Hammond, and L. Birnbaum. 2007. "TagAssist: Automatic Tag Suggestion for Blog Posts". Northwestern University. Evanston, IL, USA. http://www.icwsm.org/papers/2--Sood-Owsley-Hammond-Birnbaum.pdf [Last accessed: 5 April 2011]

K. Frantziy, S. Ananiadou, and H. Mimaz. 2000. "Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method", International Journal on Digital Libraries , 3 (2), pp. 117-132.

R. Hussey, S. Williams, and R. Mitchell. 2011. "Keyphrase Extraction by Synonym Analysis of n-grams for E-Journal Classification", Proceedings of eKNOW, The Third International Conference on Information, Process, and Knowledge Management. Pp. 83-86 (2011). http://www.thinkmind.org/index.php?view=article&articleid=eknow_2011_4_30_60053 [Last accessed: 5 April 2011]