



## COMPRENDRE: CLASSIFICATION AUTOMATIQUE OU CLUSTERING ?

La classification automatique ou clustérisation de résultats est à la mode, il suffit pour s'en convaincre de visiter les moteurs de recherche Internet suivants : <http://www.killerinfo.com/> <http://www.vivisimo.com/> <http://www.exalead.com/> <http://www.wisenut.com/> <http://www.teoma.com/> <http://www.nlsearch.com/>

Au-delà de cette vogue, cette fonctionnalité offre-t-elle un réel atout pour optimiser l'accès à l'information ? Tout d'abord, il s'agit de structurer l'information pour y accéder plus facilement. Cette organisation du document était présente bien avant l'ère des moteurs de recherche et même de l'information électronique (rangement des bibliothèques par exemple). Longtemps les recherches électroniques se sont basées sur des plans de classements appelés aussi catégories ou taxinomies, souvent créées par des documentalistes. Ces catégories étaient définies au préalable et tout document était rattaché à une ou plusieurs catégories.

Logiquement, Internet puis les Intranets ont fait exploser les volumes disponibles. Dans ce contexte, la maintenance de ces catégories est devenue d'une complexité proportionnelle à la masse des données prises en compte.

La classification automatique permet de s'affranchir de la gestion des catégories car elles sont définies automatiquement. Le procédé, s'il n'est pas nouveau, répond aujourd'hui correctement à un besoin clairement identifié, à savoir filtrer par thèmes une liste de résultats trop importante. Ce système a bénéficié des avancées technologiques autour du text mining ou extraction de connaissance à partir de textes.

Le fonctionnement est le suivant : il s'agit, pour un corpus de documents, d'analyser leur contenu et de les regrouper au sein de différents thèmes extraits. L'outil peut utiliser plusieurs algorithmes pour extraire les termes ou groupes de termes les plus fréquents :

- Un calcul de la fréquence de co-occurrences : termes très souvent proches
- Une analyse linguistique : gestion des pluriels et autres formes des termes, reconnaissance de mots composés, analyse sémantique
- Extraction de termes identifiés : nom de personne, nom d'organisation, date, lieu

Le système classe et regroupe un ensemble de documents par les expressions ainsi extraites appelées aussi « thèmes » ou « concepts ». L'ensemble de ces thèmes ou concepts est souvent appelé un « cluster ».

Les différentes solutions existantes se distinguent par les technologies utilisées et par la base de départ prise en compte pour définir les thèmes : certains prennent en compte le même ensemble de documents que celui qui est donné en résultat (répondant à des critères par exemple de résultat d'une recherche ou de nouveaux documents du jour), tandis que d'autres prennent en compte l'ensemble des documents disponibles. Le premier comportement est plus contextuel mais nécessite une technologie très rapide. Les résultats obtenus sont réellement différents mais restent pertinents selon le contexte d'utilisation. De même, selon l'application visée, la gestion de la présence de plusieurs langues ou une analyse linguistique fine peut être pertinente. Le choix doit permettre de visualiser des thèmes correspondant aux attentes des utilisateurs et de répondre notamment au critère de temps de réponse.

Ces outils paraissent promis à un bel avenir tant ils affranchissent l'utilisateur de préciser sa requête avant de la poser, en permettant néanmoins un accès à une information pertinente avec un minimum d'action.

Eric Debonne - Consultant - Solutions d'accès à l'information [www.solaci.com](http://www.solaci.com)