**Commercialisé en France par la société OIS le logiciel InfoCodex a remporté le prix de la Veille et de l'Intelligence Economique de l'édition 2006 de i-expo.
Cette solution de veille originale fondée sur un couplage de la sémantique et des réseaux neuronaux a été conçue en Suisse par deux docteurs en physique : Paul Walti et Carlo Trugenberger.**

PAR MARIANNE DABBADIE

Infocodex, une étonnante boîte à outils qui associe la sémantique aux réseaux neuronaux

InfoCodex est un moteur de recherche d'entreprise qui embarque des fonctionnalités de veille, recherche sémantique et indexation cross linguale. C'est une étonnante boite à outils qui permet d'aller interroger toutes sortes de sources, qu'elles soient en local ou sur internet, puis de trier, organiser, exploiter et croiser les résultats à des fins de veille, Business Intelligence, ou de marketing. Ce moteur permet aussi tout simplement de partager des informations catégorisées et pré- traitées entre collègues.

POUR QUELS USAGES ?

Le moteur est fourni aux clients en mode projet et les équipes d'utilisateurs sont formées selon le projet pour lequel le moteur est utilisé. Les projets sont typés de la façon suivante. Il peut s'agir de :

- Moteurs de recherche d'entreprises
- Extraction/Gestion/Capitalisation des connaissances
- Moteurs de recherche pour site Web
- Outils de veille
- Moteurs de classification et de filtrage automatique
- Indexation multilingue et cross-linguale
- Bases indexées de courriels
- Fouille de données et de textes sur des gisements de très grande dimension (Tbyte)

RÉSEAUX DE KOHONEN

La technologie d'InfoCodex est fondée sur des algorithmes sémantiques et statistiques d'une part et sur les réseaux neuronaux de Kohonen d'autre part. Les cartes auto organisées de Kohonen permettent d'extraire à partir de ces classes un ensemble de connaissances et d'informations non visibles de prime abord. Cette information auto organisée prend la forme de clusters partageant des caractéristiques communes. Le concept d'auto-organisation est d'une importance capitale dans un contexte de veille. Cela signifie que le système est capable grâce à sa finesse d'analyse, de créer des catégories auto-organisées sans apprentissage préalable.

Quelques références clients

InfoCodex compte, à titre d'exemple, parmi ses clients, L'Association Suisse de Normalisation, avec un moteur de recherche intranet/extranet comprenant un système d'alerte sur concepts pour ses publications multilingues sur les normes, Le département Recherche et Développement Siemens BT (R&D), Agroscope, avec une indexation de 15 Millions de documents et 300.000 courriels en 5 langues, ou encore Swiss Re une base de données d'antériorité et de veille sur les brevets.

...test moteur ... test moteur...test

Caractéristiques du moteur

Sur le plan technologique, le moteur InfoCodex, qui présente des capacités très avancées d'organisation et de tri des données est le résultat de la fusion de trois technologies : Linguistique, statistique et des – Réseaux de neurones auto-organisés.

Le moteur traite plusieurs langues européennes qui fonctionnent sur un mode interlingue. Il s'agit de l'Anglais, le Français, l'Allemand, l'Italien et l'Espagnol. Une requête peut ainsi être posée dans n'importe laquelle de ces langues, sur une base de données multilingue.

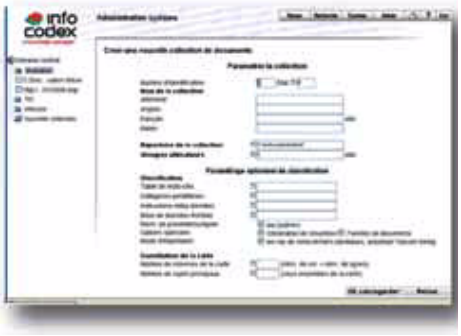
De plus, les nombreux opérateurs utilisés dans la phase d'analyse, permettent d'affiner la recherche en fonction du besoin.

L'interface est simple et facile d'utilisation, bien que sans recherche graphique particulière. Le moteur peut être lancé dans l'une des cinq langues de travail. Une interface d'administration très d'usage intuitif permet de créer, modifier, exporter des collections ainsi que de définir des métadonnées et des tables de mot-clés qui serviront lors de prochaines requêtes.

Collections de documents

InfoCodex permet de créer des collections et sous collections de documents. Pour constituer les bases de documents, il est possible d'interroger n'importe quelle source dans n'importe quel format de document. De plus, les documents écrits au format image peuvent être analysés automatiquement en OCR pour en extraire le contenu.

L'interface d'administration du système permet de créer une collection de documents en y incluant des paramètres linguistiques, des tables de mot-clés prédéfinies pour guider la recherche, demander la génération automatique de résumés, de familles de documents. En bref, toutes les fonctionnalités de prétraitement et de catégorisation peuvent être incluses dans la base dès sa création, ce qui fait gagner un temps considérable à l'analyse.

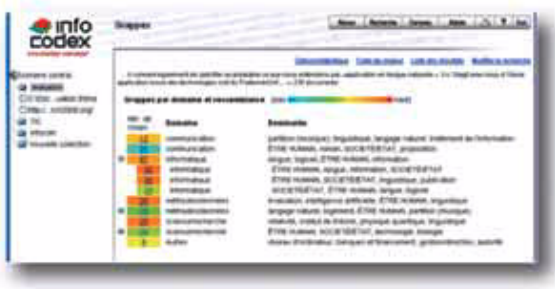


Exploiter les données d'une collection

La recherche d'information peut s'effectuer dans une collection de différentes façons. Il peut s'agir d'une recherche exacte, par synonymes ou bien par similarité. Le mode de recherche qui nous a semblé le plus intéressant sur le plan des usages de la veille est celui de la recherche par similarité.

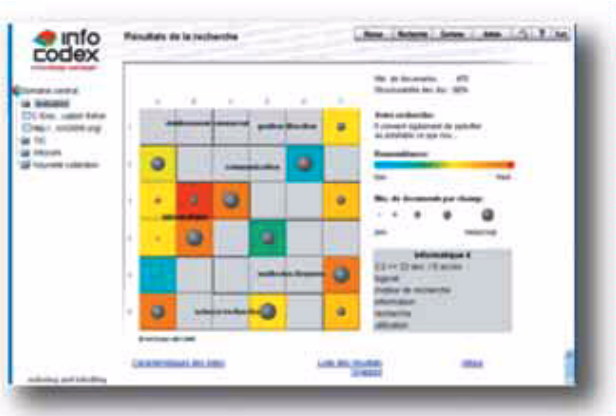
Ce type de recherche permet de retrouver de façon transversale des thématiques communes entre les documents d'une collection. Il fournit une pré-catégorisation selon les thèmes recherchés. Et ce qui est le plus surprenant c'est que le système ne se contente pas d'une requête d'une phrase. Il est capable d'analyser un document entier et de retrouver les documents qui lui sont thématiquement proches dans la base.

Dans l'exemple suivant, nous avons entré un texte d'environ une page sur l'évaluation des moteurs de recherche et effectué une recherche par similarité sur une base de données ayant pour thème la linguistique et l'intelligence économique.



Le nombre de documents par sous catégorie est affichée dans une étiquette colorée. Un code de couleur permet d'identifier automatiquement les catégories les plus pertinentes. En cliquant sur le nombre de documents, il est possible d'en afficher la liste, le résumé et aussi le contenu avec les mots de la requête et leurs synonymes mis en surbrillance.

Il est également possible de générer une "carte de chaleur" des thèmes principaux qui indique leur proximité avec la requête ou le texte initial.



La carte de chaleur ainsi générée permet de voir immédiatement quelles sont les catégories les plus pertinentes. A l'approche de la souris, les mot-clés de chaque catégorie s'affichent. Un lien au bas de l'écran permet d'afficher les textes. Dans l'exemple choisi, le texte de la recherche par similarité et certains textes de la base en anglais ont été analysés par le système et inclus dans la liste de résultats.

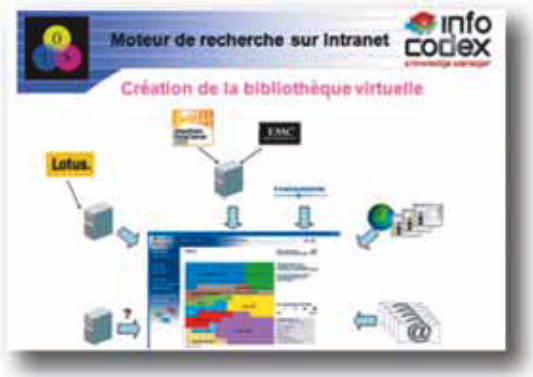
La mémoire du système

En plus de la gestion des métadonnées (création de champs de type <auteur> <marque> etc.) le système permet de créer des tables de mot-clés qui seront utilisées lors de la constitution d'une collection.

Ces tables de mot-clés peuvent être générées à partir d'une collection existante. Chaque mot-clé est alors trié manuellement par l'utilisateur en fonction de son importance. Les tables sont stockées, puis utilisées lors de l'interrogation d'une base d'information non structurée à des fins scientifiques ou professionnelles. De même, les recherches peuvent êtres stockées pour êtres re-exploitées lors de la rédaction d'un rapport ou dans le cadre d'un projet d'entreprise. Cette fonction de mémoire du système à un niveau de détail élevé et déterminé par l'utilisateur est d'une grande utilité dans les applications professionnelles.

Partage de l'information

La richesse des fonctionnalités d'InfoCodex permet de générer et de partager des bibliothèques thématiques virtuelles à des fins professionnelles.



Il va sans dire qu'avec une telle richesse de fonctionnalités créatrice de valeur grâce à la structuration d'une information provenant d'un nombre important de sources, InfoCodex, outil original de traitement de l'information non structurée, a de beaux jours devant lui.

MARIANNE DABBADIE