INSTITUT DE L'INFORMATION SCIENTIFIQUE
ET TECHNIQUE

Département Portail et Services d'information — Service Veille

Méthodologie d'évaluation d'outils de veille

Valérie Devaux – Solveig Vidal
Février 2006

Sommaire

Contexte

Pour faire face à la quantité croissante d'information, de nombreux d'outils pour la collecter, l'analyser et la diffuser ont été développés : les utilisateurs ont besoin d'aide pour les positionner sur le marché et surtout pour effectuer leur choix.
Il existe des études comparatives mais elles présentent des limites : elles sont statiques, proposent peu de critères, ou des listes de critères plus détaillées et sont alors réalisées par un éditeur de logiciel. Seule l'étude menée par Yasmina Quatrain & al propose une méthodologie détaillée[1].
D'autres raisons ont amené le service veille de l'INIST-CNRS à proposer ce benchmarking :

- ✓ en tant que producteur et utilisateur d'outils de veille, il effectue une veille pour ses propres besoins ;
- ✓ son équipe est constituée d'ingénieurs aux compétences reconnues ;
- ✓ en tant qu'unité du CNRS, son indépendance est à souligner.

Objectifs

Ce projet a 3 principaux objectifs :

1. **Recenser et décrire les outils de veille** présents sur le marché ou des prototypes
   Le recensement permettra de constituer un annuaire des outils intervenant dans le cycle de veille : les outils de collecte, d'analyse et de diffusion de l'information. Intéressée par ce projet, la section Recherche de l'ADBS, dont plusieurs professionnels de l'information adhérents se trouvent confrontés au choix d'un outil, nous a demandé de concentrer nos efforts, dans un premier temps, sur les outils d'analyse.

2. **Proposer une méthodologie et des tests en ligne**
   Nous voulons aller au-delà d'une fiche présentant les fonctionnalités de ces outils en élaborant des tests. Afin d'y parvenir, des questions type que se posent les veilleurs ont été élaborées pour évaluer les outils d'analyse. Tous les résultats sont accessibles sur internet : http://outils.veille.inist.fr

3. **S'ouvrir en transférant notre démarche d'évaluation d'outils vers une collaboration.**
   Vu l'ampleur et l'ambition du projet (nombre d'outils, mise à jour des versions), l'aspect collaboratif est pratiquement incontournable que ce soit au niveau :
   - ✓ des outils (alimenter le recensement des outils, corriger/ajuster les fiches outils du recensement, participer aux tests)
   - ✓ de la méthodologie (amélioration de la méthodologie, propositions de nouveaux tests, veille sur les procédures de tests)

   Voulant impulser le mouvement en établissant d'abord une méthodologie, cet aspect collaboratif, bien que fondamental, viendra dans un second temps.

[1] JADT 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles : Évaluation d'outils de Text Mining : démarche et résultats. Yasmina Quatrain, Sylvaine Nugier, Anne Peradotto, Damien Garrouste. http://www.cavi.univ-paris3.fr/lexicometrica/ jadt/jadt2004/pdf/JADT_089.pdf

Méthodologie

Fiche descriptive

Nous nous sommes appuyés sur différentes études[2] pour constituer cette fiche composée actuellement de 70 éléments enregistrés dans une base de données SQL à partir d'un formulaire PHP[3]. Pour des raisons d'objectifs (nombre important d'outils à tester, assurer la pérennité de notre méthode sans pour autant la figer de manière définitive) mais également de temps et de moyens, nous avons volontairement limité le nombre de critères. Sont pris en compte :

- ✓ les éléments d'ordre commercial (identité du logiciel, tarif, aide en ligne, support technique) ;
- ✓ les éléments d'ordre technique (pré-requis technique, modalités d'installation, d'administration, ergonomie de l'interface, formats des documents) ;
- ✓ les fonctionnalités liées à la requête, à la collecte, à l'analyse, à la diffusion.

Cette fiche vise à donner un aperçu global de l'outil et devrait permettre à l'utilisateur de faire une première sélection en fonction de ses besoins quand, sur le site, les possibilités de recherche multicritère seront implémentées (Annexe 1 : fiche descriptive).

Evaluation des outils d'analyse : « Corpus test » et grille d'évaluation

Corpus de données structurées – notices bibliographiques PASCAL

Afin de tester ces outils, différents corpus ont été constitués.
Producteur des bases de données bibliographiques PASCAL et FRANCIS, l'INIST-CNRS, et plus particulièrement le Service Veille, maîtrise ces données structurées en format XML. Cette raison explique le choix de constituer, dans un premier temps, des corpus issus de PASCAL. Le choix de la thématique s'est porté sur les OGM : sujet à l'origine de nombreuses publications et pour lequel nous avons également les compétences scientifiques en interne.
Il nous a semblé intéressant de tester les outils en fonction de la taille du corpus, d'où 3 corpus de respectivement 892, 5000 et 10319 notices. Des tests sur des corpus de taille encore plus conséquente, de l'ordre de 50000 à 100000 références, seront effectués.
Par la suite, seront testés des corpus constitués à partir de plusieurs bases de données bibliographiques, afin de mettre en avant les éventuelles possibilités de dédoublonnage des outils, puis des corpus comportant du texte intégral se présentant sous différents formats afin d'aborder les capacités de traitements sémantiques.

---

[2]
- ✓ Benchmark : Solutions de veille stratégique. Digimind. Janvier 2004. Accessible depuis http://www.digimind.com
- ✓ CIFT 2004 : Evaluation d'outils de text mining dans un contexte industriel. Yasmina Quatrain, Sylvaine Nugier, Anne Peradotto http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/12/56/sic_00001256_00/sic_00001256.pdf
- ✓ Guide pratique : veille et gestion des connaissances. Mars 2005

[3] Pour chacun de ses 70 éléments, une zone de commentaire est proposée au testeur. Une aide en ligne guide ce dernier dans l'alimentation de cet espace.

Questionnaire

Près de 50 questions tentent de cerner les principales interrogations que se posent les veilleurs lors de leurs études (Annexe 2). Il faut souligner que ce premier lot est lié à des données structurées. Leur organisation est liée à des problématiques de recherche et non pas aux fonctionnalités de l'outil[4]. Les questions concernent :

- ✓ le corpus dans sa globalité ;
- ✓ une thématique spécifique au sein du corpus ;
- ✓ un auteur ;
- ✓ un organisme ;
- ✓ un pays.

Dans un premier temps, la question est de savoir si, à l'aide de l'outil, on peut répondre à ces différents points. Quatre réponses sont alors possibles :

- ✓ « Oui » lorsque l'information est obtenue sans difficulté.
- ✓ « Oui mais difficilement » quand l'information est obtenue suite à plusieurs échecs ou essais (au moins deux essais), ou en ayant dû consulter une aide (en ligne ou pas).
- ✓ « Non » en cas d'échec suite soit à une anomalie du logiciel, soit à une difficulté d'utilisation trop importante ayant mis le testeur en échec.
- ✓ « Sans objet » quand le logiciel n'est pas initialement conçu pour répondre à ce besoin.

Une zone de commentaire pour chaque groupe de questions permet d'apporter des éclaircissements particuliers, nuancer des réponses parfois trop abruptes.

Une prochaine étape est l'intégration de la notion de pertinence des résultats. Cette notion de référentiel est possible puisque nous connaissons de manière approfondie les corpus explorés à l'aide d'outils conçus et utilisés par le Service Veille de l'INIST dédiés au traitement de nos données.

« Notation »

L'attribution d'une note est un aspect délicat. Le résultat obtenu ne peut pas toujours être assimilable à la capacité globale de l'outil, la note étant sensible à plusieurs paramètres : le système de notation lui-même, le testeur, le corpus, le questionnaire.
Ainsi nous sommes nous pour le moment arrêtés à une comptabilisation des types de réponse. Ultérieurement, nous présenterons de manière visuelle, à l'aide d'« aires fonctionnelles », les résultats des tests en fonction de critères[5].

Choix des logiciels

La sélection des premiers logiciels d'analyse testés s'est faite en fonction de la disponibilité simple et rapide de ces outils. Nous avons donc choisi ceux que nous détenions en interne : KeyWatch (i-Scope), LexiquestMine (SPSS). Nous avons le plus tôt possible élargi notre cercle et Intellixir a accepté de participer aux tests.

[4] Par exemple, les questions « Quels sont les pays associés au terme "sécurité alimentaire" ? » et « Quelles sont les thématiques étudiées ? (ici par la France) » ne sont pas dans la même catégorie de questions alors qu'elles sont liées à la même fonctionnalité (la cooccurrence interchamp : pays / mots-clés)
[5] CIFT 2004 : Evaluation d'outils de text mining dans un contexte industriel. Yasmina Quatrain, Sylvaine Nugier, Anne Peradotto http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/12/56/sic_00001256_00/sic_00001256.pdf (p.10)

Dans l'objectif de constitution d'un annuaire, le nombre d'outils doit évoluer. En fonction des contacts avec les éditeurs et d'autres organismes scientifiques, nous alimenterons régulièrement le site au rythme d'un nouvel outil chaque mois à partir du printemps 2006.

Perspectives

A chaque étape du cycle de veille, correspond un outil spécifique donc des tests différents.

Outils

- ✓ **Outils de collecte**

  Pour les outils de collecte, est en cours de réalisation un « crash site », site où l'on maîtrisera la chronologie des modifications du site, de ses pages (ajouts - suppressions), la complexité technique (pages statiques et dynamiques, la profondeur des liens …), les types de documents …

- ✓ **Outils d'analyse**

  Les prochaines étapes concerneront l'élargissement du panel des outils testés, la constitution des corpus à partir de notices bibliographiques issues de plusieurs bases de données, puis de données en texte intégral. Ces points nécessiteront de réaliser de nouveaux questionnaires, ou de compléter le questionnaire existant.

- ✓ **Outils de diffusion**

  Cet axe commencera une fois que les méthodologies concernant les outils précédents seront établies.

- ✓ **Priorités**

  Nous donnons actuellement la priorité aux outils d'analyse de données structurées et seuls les tests les concernant sont opérationnels, viendront ensuite les outils de collecte, puis les outils d'analyse de documents en texte intégral.

Travail collaboratif – Groupe de travail

Devant l'ampleur du projet, l'aspect collaboratif est incontournable. Nous avons initié l'ouverture en mettant à disposition une méthodologie, un outil permettant la diffusion d'une étude dynamique (contrairement aux études statiques actuelles) de comparaison d'outils de veille.
Cette première phase a permis d'identifier les principales fonctionnalités utiles aux veilleurs. Afin d'être plus exhaustif et de satisfaire l'ensemble de la communauté concernée (utilisateurs-éditeurs), des échanges sont à mettre en place.
La gestion du site à l'aide de SPIP sera un premier pas vers un travail collaboratif. Par ailleurs, afin de ne pas directement et entièrement dépendre des éditeurs, est envisagé un appel à participation auprès des membres de l'ADBS et des documentalistes de plusieurs EPST (INSERM, INRA, INRIA) ou EPIC (CNES), qui détiendraient des outils de veille dans leur organisme.

Relations avec les éditeurs

Les relations avec les éditeurs sont pour nous fondamentales à plusieurs titres :

- ✓ la mise à disposition de l'outil ou un accès aux résultats du traitement des données pour réaliser les tests ;

- ✓ un complément de notre suivi de l'actualité et des tendances qui constituera une animation sur le site.

Pour les éditeurs et prestataires de services, ce projet se présente comme une vitrine importante : présentation à des clients potentiels, mais également retour formalisé sur leurs produits.
Afin de maintenir l'indépendance, des relations claires avec les éditeurs doivent être établies : les tests suivent une procédure bien définie ; suite à la diffusion des résultats les éditeurs ont un droit de réponse, notamment dans le cas d'outils présentant des qualités ou objectifs spécifiques n'apparaissant pas forcément dans un projet à vocation plus généraliste.

Conclusion

Afin de répondre au mieux aux besoins de l'utilisateur dans la sélection d'un outil de veille, nous proposons de référencer, évaluer et comparer, suivant un mode dynamique, le plus d'outils possible à l'aide d'une méthodologie souhaitée la plus aboutie.
Pour y parvenir et satisfaire la communauté concernée, un travail collaboratif est nécessaire entre utilisateurs, utilisateurs potentiels, éditeurs et prestataires de services.

Annexe 1 : Fiche descriptive

En italique les commentaires pour aider à remplir la fiche

Identité
Nom du logiciel
Date d'enregistrement :

Test du logiciel :
- Logiciel testé
- Informations recueillies

Etat d'avancement de l'outil
- Prototype
- Logiciel commercialisé
- Logiciel libre

Type d'outil
- Collecte
- Analyse
- Intégré

Version testée

Nature de la version testée
- Logiciel acheté
- Logiciel téléchargé gratuitement
- Service ASP
- Logiciel prêté (cd-rom)
- Sans objet
- Numéro ou nom de version
- Url
- Durée d'accès ou utilisation (en jours)
- Comparatif des versions :
  Donner les avantages de la version testée ou recensée par rapport aux versions antérieures. Préciser également la périodicité des mises à jour et l'évolutivité.

Nom de la société
- Adresse
- Téléphone
- Mail
- Url

Descriptif du logiciel
Descriptif concis mettant en avant les principales fonctionnalités de l'outil.

Descriptif commercial

Oui : *Le descriptif commercial est réservé aux outils qui ne sont pas encore testés.*
Non

Généralités

Prix

Gratuit
Non communiqué
Moins de 1000 €
Entre 1000 et 5000 €
Entre 5000 et 10000 €
Plus de 10000 €
Prix exact : *Mentionner l'aspect abonnement si besoin*

Prix Mono/Multi

Monoposte
Multiposte

Aide en ligne

Aucune
Une démonstration en ligne
Un manuel

Support Technique

Oui
Non
Pas d'information
Préciser coût de l'assistance + formation

Pré-requis technique

Indiquez le système d'exploitation, mode client/serveurs, les ressources mémoires, l'espace disque nécessaire, le(s) logiciel(s) nécessaire(s), le SGBD (Oracle, MySQL...) et le mode web

Multilinguisme

Oui
Non
Langues
- Français
- Anglais
- Allemand
- Espagnol
- Italien
- Autres

Cross language

Oui
Non

Reconnaissance automatique

Oui
Non
Autre information sur la gestion du multilinguisme

Procédure d'installation

Installation assistée sur monoposte
Nécessite un informaticien (monoposte)
Installation en mode multiposte
Logiciel non testé
Autre information

Paramétrage en tant qu'administrateur

Facile
Modérée
Difficile
Sans objet
Logiciel non testé
Autre information

Modéré = familiarisation du produit nécessaire. Difficile = formation indispensable

Utilisation, en tant qu'utilisateur final

Facile
Modérée
Difficile Logiciel non testé
Autre information

Modéré = familiarisation du produit nécessaire. Difficile = formation indispensable

Informations sur l'interface Homme/Machine

Personnalisation (ex. couleurs-modules)

Fonctionnalités

Type de documents traités

Structuré (Référence bibliographique)
Non structuré (Texte intégral)
Autre information :
Préciser si besoin d'un format particulier notamment dans les statistiques sur données structurées.
Préciser la taille maximum éventuelle des noms de variables

Formats des documents traités

TXT
HTML
XML
PDF
Microsoft Office
RSS
Autre information
Préciser si besoin de « prénormalisation » des documents (orthographe, harmonisation, majuscules)
Préciser s'il faut ou non dédoublonner en amont

Rédaction de requêtes

Taille de la requête
- < 5 mots
5 - 10 mots
> 10 mots
Taille en nombre de caractères
Utilisation d'opérateurs
Proximité
ET OU SAUF
Troncature à droite
Troncature à gauche
Parenthèses
Expression exacte
Autre information : *bien distinguer les modalités de requête pour crawler, pour rechercher dans un corpus, recherche sur texte intégral, recherche multicritère (champs)*

Fonctionnalités liées à la collecte (ce module évoluera ultérieurement)

Moteur de recherche
Métamoteur
Surveillance d'une page
Détection d'une nouvelle page
Détection d'une page supprimée
Crawl d'un site
Crawl d'une partie d'un site
Crawl d'une Url individuelle
Crawl d'une liste d'Urls
Crawl via une requête
Crawl d'Urls extérieures au site initial
Possibilité de paramétrer la profondeur du crawl
Possibilité de paramétrer le sens du crawl
Autre information

Fonctionnalités liées à l'analyse statistique

Pour la référence bibliographique
- Comptage d'occurrence (Intrachamp)
- Comptage de cooccurrence (Intrachamp)
- Comptage d'occurrence (Interchamp)

Pour le document en texte intégral
- Comptage d'occurrence
- Comptage de cooccurrence

Autre information
Préciser si l'hapax est calculé (occurrence où un auteur-organisme-pays apparaît seul)
Préciser si on peut paramétrer- pondérer les fréquences)

Fonctionnalités liées à l'analyse linguistique

Analyse morpho-lexicale
Analyse syntaxique
Analyse sémantique
Extraction terminologique
Gestion de dictionnaires
Recherche d'entités nommées
Autre information :
Préciser si dans l'analyse morpho-lexicale on est en présence de lemmatisation, stemmatisation …
Préciser si dans l'analyse syntaxique on est en présence d'une reconnaissance du rôle grammatical dans la phrase, d'une reconnaissance des groupes nominaux

Catégorisation

Oui
Non
Possibilité de mise à jour
Possibilité d'apprentissage
Méthode employée

Classification

Oui
Non
classification hiérarchique
Non Hiérarchique
Possibilité de mise à jour
Méthode employée

Fonctionnalités liées à la diffusion

Envoi de résultats

Alerte(s) par email
Autre information

Présentation graphique

Listes
Tableaux
Cartographie
Courbes de tendances/évolution
Diagrammes de répartition
Autre information

Formats d'exports

TXT
HTML
XML
PDF
Microsoft Office
RSS
Autre information

Accès aux documents

Oui
Non

Commentaire du testeur

Annexe 2 : 52 questions « Corpus test »

Logiciel et corpus d'étude
Logiciel d'analyse testé ?
Version du logiciel testé ?
Volumétrie du corpus d'étude (en nombre de documents) ?
Nom du testeur :
Date (jj-mm-AAAA) :

Questions générales (Pouvez-vous répondre à ces questions ?)

- ✓ Quels sont les auteurs les plus productifs ?
- ✓ Quels sont les réseaux d'auteurs liés à cette thématique ?
- ✓ Quels sont les réseaux d'auteurs les plus productifs ?
- ✓ Quels sont les organismes les plus productifs?
- ✓ Quels sont les pays les plus productifs ?
- ✓ Quelle est la répartition par année de l'ensemble des documents du corpus ?
- ✓ Quels sont les nouveaux concepts ?
- ✓ Quels sont les nouveaux auteurs ?
- ✓ Quels sont les nouveaux organismes ?
- ✓ Quels sont les nouveaux pays ?
- ✓ Informations supplémentaires

Questions relatives à un thème plus précis (Pouvez-vous répondre à ces questions ?)

- ✓ Combien d'articles sont consacrés à la "sécurité alimentaire" ?
- ✓ Quels sont les organismes associés au terme "sécurité alimentaire" ?
- ✓ Quels sont les auteurs associés au terme "sécurité alimentaire" ?
- ✓ Quels sont les réseaux d'auteurs qui travaillent sur ce sujet ?
- ✓ Quels sont les réseaux d'auteurs les plus actifs qui travaillent sur ce sujet ?
- ✓ Quels sont les pays associés au terme "sécurité alimentaire" ?
- ✓ Quel est le nombre de publications par année lié au terme "sécurité alimentaire" ?
- ✓ Quels sont les nouveaux concepts liés au terme "sécurité alimentaire" ?
- ✓ Quels sont les nouveaux auteurs liés au terme "sécurité alimentaire" ?
- ✓ Quels sont les nouveaux organismes liés au terme "sécurité alimentaire" ?
- ✓ Quels sont les nouveaux pays liés au terme "sécurité alimentaire" ?
- ✓ Informations supplémentaires

Questions relatives à un auteur (Pouvez-vous répondre à ces questions ?)

- ✓ Combien d'articles a écrit STRAZIELLE (C.) ?
- ✓ Sur quelles thématiques travaille-t-elle ?
- ✓ Pour quel organisme travaille-t-elle ?
- ✓ Avec quel(s) organisme(s) travaille-t-elle ?
- ✓ Avec quel(s) auteur(s) travaille-t-elle ?
- ✓ Quelle est la nationalité des organismes avec lesquels(s) elle travaille ?
- ✓ Dans quel pays travaille-t-elle ?
- ✓ Quel est le nombre de publications par année pour STRAZIELLE (C.) ?
- ✓ Quels sont les nouveaux concepts associés à STRAZIELLE (C.) ?
- ✓ Quels sont les nouveaux auteurs associés à STRAZIELLE (C.) ?
- ✓ Quels sont les nouveaux organismes associés à STRAZIELLE (C.) ?
- ✓ Quels sont les nouveaux pays associés à STRAZIELLE (C.) ?
- ✓ Informations supplémentaires

Questions relatives à un organisme (Pouvez-vous répondre à ces questions ?)

- ✓ Quel(s) laboratoires(s) CNRS travaille(nt) sur les OGM ?
- ✓ Sur quel sujet travaille l'UMR 5004 ?
- ✓ Avec quel(s) pays travaille l'UMR 5004 ?
- ✓ Quels sont les chercheurs qui travaillent avec ou pour l'UMR 5004 ?
- ✓ Quel est le nombre de publications par année de l'UMR 5004 ?
- ✓ Quels sont les nouveaux concepts associés à l'UMR 5004 ?
- ✓ Quels sont les nouveaux auteurs associés à l'UMR 5004 ?
- ✓ Quels sont les nouveaux organismes affiliés à l'UMR 5004 ?
- ✓ Quels sont les nouveaux pays associés à l'UMR 5004 ?
- ✓ Informations supplémentaires

Questions relatives à un pays (Pouvez-vous répondre à ces questions ?)

- ✓ Quel est le nombre d'articles publiés en France ?
- ✓ Quelles sont les thématiques étudiées ?
- ✓ Quels sont les auteurs français ?
- ✓ Quels sont les auteurs (français et étrangers) liés à la France ?
- ✓ Quels sont les organismes français ?
- ✓ Quels sont les organismes (français et étrangers) liés à la France ?
- ✓ Quels sont les pays qui travaillent avec la France ?
- ✓ Quel est le nombre de publications par année pour la France ?
- ✓ Quels sont les nouveaux concepts liés à la France ?
- ✓ Quels sont les nouveaux auteurs (français et étrangers) liés à la France ?
- ✓ Quels sont les nouveaux organismes (français et étrangers) liés à la France ?
- ✓ Informations supplémentaires

Commentaire du testeur