ANNEXE VII.1

Proposition et discussion de protocoles d'évaluation pour une application de diffusion ciblée

Notes de travail

Table des matières de l'Annexe VII.1

1. Remarques : difficultés particulières

- A la différence d'un classement, le nombre de bonnes réponses est indéfini. Deux propositions de liste de diffusion sont le plus souvent incommensurables (indécision sur la valeur des nouveaux rapprochements).
- Besoin d'un corpus large (représentatif en fait), et peut-être aussi renouvelé ; il s'agit d'éviter que le corpus-test se transforme en corpus d'apprentissage (au sens négatif : vue restreinte et figée).
- Recueillir des données sur les recherches documentaires, les emprunts et les commandes d'ouvrages, les abonnements, voire sur les documents effectivement utilisés, les fiches de lecture réalisées, serait très difficile à suivre, et serait intrusif.

2. Voies envisageables

a) Evaluation interne, par utilisation d'informations parallèles

Pour certains types de documents, élaborés ou/et bien connus à la DER, on peut disposer d'informations codées sous formes de liens inter-documents, ou dans des rubriques structurées (grilles fixes), et qui ne sont pas exploitées (ou en tout cas pas exploitées pour elles-mêmes) dans le calcul des rapprochements. Typiquement : l'auteur d'une Note interne, les versions successives d'un même document, les articles commandés par un même agent, le rattachement à un même projet (ARD/AID, PPRD), l'émission par un agent (ou un collectif) relevant de tel Groupe, tel Département et tel Service, l'existence de liens contractuels (partenariat avec un laboratoire par CERD par exemple), etc.

Exploiter de tels renseignements fournit un mode de contrôle à la fois précis, bien automatisable, et fiable. Cependant :

- cela n'est possible que sur un sous-ensemble de documents, sous-ensemble non nécessairement représentatif.
- les informations disponibles dépendent des corpus : il faut procéder au cas par cas.
- un document atypique (et par exemple relatif à un concept en émergence, à une problématique nouvelle -avec dans ce cas une importance stratégique évidente-) sera généralement mal caractérisé à travers les grilles standardisées et statiques : les indications sont valables dans le cas général mais peu adéquates pour positionner des éléments isolés. Or la puissance de calcul et la mémoire maintenant disponibles doivent justement permettre de traiter finement des cas que l'on était auparavant tenu de négliger.

Notons que c'est notamment pour éviter des limitations de ce genre que la diffusion ciblée tire l'essentiel de ses données du texte et de son analyse linguistique sémantique.

b) Evaluation en aval, au moyen d'un échantillon

Un autre mode connu de validation fait appel aux statistiques. Il requiert la donnée (ou le recueil et la constitution) d'un échantillon représentatif de résultats avec un code d'appréciation renseigné par l'utilisateur.

Les principaux avantages de ce système tiennent au crédit fait aux opinions des utilisateurs (on se donne le moyen de les prendre en compte de façon directe), et au pouvoir expressif des chiffres obtenus (mesure des progrès, compétition de différentes méthodes,...).

Ses limitations viennent d'une part de son coût (l'effort pour constituer un tel corpus de validation est très grand, et l'on consomme aussi une part non négligeable du temps des utilisateurs), d'autre part du décalage induit sur plusieurs plans :

- décalage temporel : on a une "photographie" d'une population donnée à une date donnée. Sauf extrapolations (mais alors on perd en fiabilité et en qualité de l'information), ces données de contrôle sont statiques.
- décalage créé par les conditions de l'expérience : le recueil des résultats se fait toujours dans un cadre, un contexte particulier, non complètement reproductible. La présentation de l'expérience, la formulation des questions, la grille de choix permis pour les réponses, etc., ne sont bien sûr pas sans impact.

- décalage induit par l'expression *a posteriori* des utilisateurs, sur les résultats en aval du système : le calcul des proximités d'une part, et le jugement émis d'autre part, sont dissociés dans le temps. Il en résulte que l'opinion exprimée est conditionnée, à des degrés variables, par l'image que l'on se fait du système, et par ce que le système "laisse voir" : par exemple, on reste très démuni pour mesurer le silence (le système reste muet sur les documents dont il n'a pas saisi la pertinence du rapprochement, il ne fait aucune suggestion de rattrapage).

Ceci sans poser le classique mais épineux problème de la représentativité de l'échantillon.

Une validation de ce type (analyse d'un échantillon de réponses) a été réalisée par (Vavasseur & Lemesle 1994) : elle est riche d'enseignements quant à la qualité du « Qui Fait Quoi ? 94 ».

(Sta 1994) a aussi conduit une étude statistique poussée sur les résultats recueillis pour le « Qui Fait Quoi ? 93 ». Il en tire une conclusion opérationnelle : 9 descripteurs générant du bruit ont été isolés automatiquement, et le gain de précision apporté par leur suppression est évalué à 6 points.

c) Evaluation directe, avec des utilisateurs

Un exemple de ce type de validation est donné dans notre cas par les expériences-pilotes conduites par Laurent Vavasseur en décembre 1994. Il s'agit, en présence d'un destinataire, d'évaluer en direct l'acceptabilité des rapprochements calculés (envois proposés), et d'étudier l'impact de corrections et d'ajustements manuels du profil.

Il va sans dire que de tels rendez-vous, préparés et animés avec soin, demandent un investissement conséquent (mobilisation forte d'au moins un expert en Diffusion Ciblée, ayant une bonne connaissance du système). Ils apportent quantitativement un nombre d'information limité, mais qualitativement permettent de saisir des indications précieuses. Si l'on perd en systématicité des résultats, on gagne en richesse avec la potentialité d'avoir des renseignements imprévus.

La grande force de ce type de méthode réside sans doute dans la coopération entre deux expertises : celle du chercheur/administrateur de l'outil de Diffusion Ciblée, qui a une connaissance fine et juste du système, et celle de l'ingénieur-chercheur, spécialiste de son domaine, et conscient de l'utilité concrète d'un document, "sur le terrain".

La validation directe avec des utilisateurs a donc plus un caractère suggestif que conclusif : plutôt que d'enregistrer les performances du système, elle nourrit et oriente les recherches en vue d'améliorer le système (sur le plan de la technique, du service offert, etc.).

d) Evaluation par suffrage (préférences des utilisateurs)

Pour quelques choix de réalisation que l'on veut valider, le principe consiste à présenter les alternatives possibles comme différents paramètres, et à observer lesquelles les utilisateurs adoptent dans la pratique.

L'avantage de cette manière de faire est d'avoir des informations renseignant sur la pratique effective des utilisateurs, dans des situations concrètes, sans que cela soit intrusif. De plus, il est techniquement facile de recueillir les choix de paramétrage des traitements.

Cette approche présente néanmoins des limites. Elle ne convient que pour comparer seulement deux ou trois possibilités aux différences assez sensibles. C'est une évaluation lente, qui demande une certaine durée (plusieurs mois). Surtout, on est conduit à concevoir les évolutions du système en termes d'opposition et de concurrence plutôt que de composition. Enfin, obtenir des suffrages significatifs suppose des utilisateurs suffisamment motivés et disponibles pour être curieux d'explorer les diverses possibilités du système et avoir la patience de les expérimenter. On notera encore que le point de vue est rétrospectif et non prospectif (notamment en ce qui concerne les corpus examinés : c'est ce qui s'est fait, non ce qui pourrait se faire).

e) Evaluation dynamique sur feed-back

C'est l'approche adoptée par certains systèmes à base de profils. Lorsque l'utilisateur interroge le système, pour chaque proposition qui lui est présentée, il lui est demandé indiquer un jugement de « pertinence ». Ces retours ont pour but de corriger et d'affiner le profil, à l'origine de la sélection.

Pour la diffusion ciblée, il faudrait revoir les modalités pour exprimer un jugement. Un certain nombre de cas sont à prévoir : un destinataire trouvé par le système mais auquel l'utilisateur a déjà pensé a une autre forme de « pertinence » qu'un destinataire inconnu mais qui semble tout a fait concerné par le sujet du document, ou encore qu'un troisième destinataire qui est moins proche du sujet mais serait intéressant à contacter pour d'autres raisons, en partie liées. En somme, il est au moins clair que *pertinence* / *non pertinence* n'est pas équivalent à *envoi (diffusion)* / *non envoi*.

Le point fort est ici la représentativité des situations d'interrogation : type d'envois, utilisateurs touchés, actualité des usages, mode d'appropriation de l'outil et d'appréciation des résultats, satisfaction exprimée.

Une telle démarche suppose des utilisateurs suffisamment motivés (du moins un « bon » ensemble d'entre eux), car l'indication systématique de jugements devient rapidement laborieuse et pesante. Une conséquence prévisible est l'hétérogénéité des retours, non seulement quant au soin apporté à l'évaluation des propositions, mais aussi quant à la signification et la valeur accordée à chaque catégorie prévue. Comme la plupart des autres modes d'évaluation, le point de vue est rétrospectif et non prospectif : on s'occupe d'usages constatés, plutôt que de déceler et de contribuer à faire apparaître des usages possibles intéressants.