Un outil pour surmonter la surcharge d'information de la veille stratégique

Annette CASAGRANDE (*), **Humbert LESCA** (*), **Laurent VUILLON** (**)
annette.casagrande@upmf-grenoble.fr, humbert.lesca@upmf-grenoble.fr, laurent.vuillon@univ-savoie.fr

(*)CERAG, Université Pierre Mendès France Grenoble BP 47 38040 Grenoble Cedex 9 France,
(**) LAMA, bâtiment Chablais, Campus Scientifique, 73376 Le Bourget-du-Lac Cedex France

Mots clefs :
Surcharge d'information, veille stratégique, informations voisines
Keywords:
Information overload, anticipatory business environment scanning, adjacent information
Palabras clave:
Sobrecarga de información, inteligencia estratégica anticipativa e informaciones relacionadas.

Résumé

Dans cette communication, nous proposons le concept d'« informations voisines », nous indiquons son utilité dans le processus de veille anticipative stratégique (VAS), face au problème de la surcharge d'information notamment occasionnée par l'usage de l'Internet. Nous présentons un prototype d'outil informatique visant à instrumenter le concept ainsi qu'un cas d'application. Le concept est particulièrement utile lorsque la veille stratégique est orientée « exploitation des informations à caractère anticipatif » pour l'anticipation, ceux-ci étant généralement noyés dans de gros volumes de données. Nous expérimentons notre prototype sur la problématique de la valorisation du CO_2 et nous montrons ainsi que cet outil permet un réel gain de temps pour rendre utilisable les informations collectées, par les décideurs.

1 Introduction

Si la surcharge d'information n'est pas un phénomène nouveau [2] , elle prend une ampleur grandissante du fait de l'utilisation toujours plus importante des nouvelles technologies de l'information dans l'entreprise, notamment Internet.

D'après la littérature, la surcharge d'information est un concept à plusieurs dimensions [10] :

- informationnelle : la surcharge est envisagée du point de vue de la quantité d'informations reçues. [7]
- communicationnelle : la surcharge d'information provient également des moyens de communications électroniques tels que le mail [16] [[7]
- cognitive : la surcharge cognitive est liée au fait que les individus ont des capacités cognitives limitées contrairement au volume d'information reçu qui ne cesse de croître [10].

L'étude IDC [8] montre que d'ici 2020 les volumes de données générées devrait être multipliés par 15. En outre, les informations les plus difficiles à gérer comme les flux texte, audio et vidéo non structurés représentent 80 % de ces données. Ceci n'est pas sans conséquences pour les salariés qui sont confrontés à un triple défi [10] :

- l'accélération de la prise de décision. La perception par les salariés de l'exigence de prendre des décisions dans un laps de temps plus court est passée de 65.5% en 2001 à 70.9% en 2005.
- l'augmentation de la surcharge informationnelle. En 2005, les salariés étaient 79.3% à considérer que le volume d'informations à traiter était trop important alors qu'ils étaient 71.7% en 2001.
- l'augmentation de la surcharge cognitive. La proportion de salariés estimant « passer davantage de temps à classer l'information » est passée de 42.9% en 2001 à 57.2% en 2005.

Or « la saturation d'informations conduit d'abord à la **dégradation du processus de décision**. Les recherches montrent qu'il existe un nombre optimal d'informations à recueillir pour prendre une décision. Au-delà d'une certaine quantité d'information, la qualité du processus décisionnel baisse, tant d'un point de vue de la qualité (décision rationnelle dans le contexte), que du temps pour prendre la décision (une décision qui intervient trop tard n'est pas bonne) » [18] .

Dans ce contexte, la veille stratégique a pour objet l'amélioration du processus décisionnel des entreprises. Cette discipline est basée sur l'observation et l'analyse de l'environnement scientifique, technique, technologique et les impacts économiques présents et futurs pour en **inférer** les menaces et les opportunités de développement [6] . L'équipe de veille stratégique du CERAG a proposé la méthode VASIC pour « permettre à l'entreprise d'agir vite, au bon moment, avec le maximum d'efficacité et le minimum de ressources, dans le but de contribuer à sa compétitivité durable » [12] . Pourtant, cette méthode, pertinente dans le cas d'informations préalablement sélectionnées, est mise en échec de par l'absence d'outils appropriés de gestion de la surcharge d'information.

Ce manque d'outils se ressent dans la phase d'interprétation des informations et notamment dans la préparation de la séance de création collective de sens qui est au centre de la méthode VASIC. On entend par création collective de sens : « un groupe de personnes [qui] accepte volontairement de mettre en commun (en collectif) leurs capacités de détecter des événements, d'en parler, de les interpréter ensemble et d'en tirer des enseignements utiles pour l'action» [12] . Pour cela, il est nécessaire de sélectionner parmi les nombreuses informations (souvent 'données brutes') collectées par l'entreprise des informations entre lesquelles il est possible d'établir des liens. On parlera alors d'« informations voisines ». Face à la surcharge d'information, comment sélectionner des informations voisines ? Comment mesurer la proximité (le voisinage) entre deux informations ? Nous porterons notre attention sur les informations numériques textuelles

Nous définirons dans une première partie le contexte de cette recherche ainsi que la problématique. Nous présenterons dans une seconde partie le prototype développé pour répondre à cette problématique et nous décrirons l'expérimentation de cet outil dans le cadre d'un projet de veille stratégique. Nous en tirons quelques enseignements dans la conclusion.

2 Contexte et problématique

2.1 Veille anticipative et Création Collective de Sens

Le processus VASIC est **défini** comme suit : « Processus collectif et proactif par lequel des membres de l'entreprise traquent (perçoivent et choisissent) de façon volontariste, et utilisent des informations à caractère anticipatif et pertinentes concernant leur environnement extérieur et les changements pouvant s'y produire » [1]. Par information à caractère anticipatif nous désignons des informations utiles pour concevoir des anticipations d'événements qui seront susceptibles d'influer sur le devenir de l'entreprise Le processus est illustré par la Figure 1.

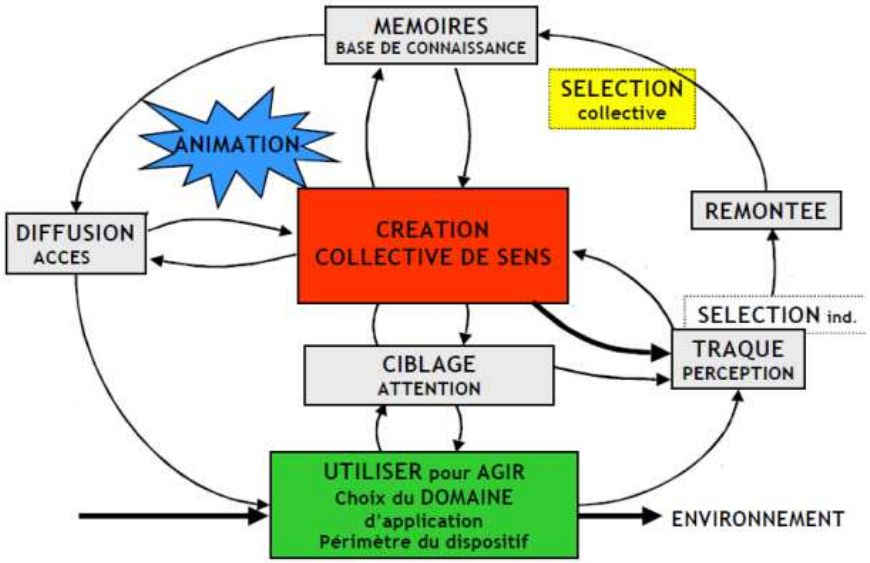


Figure 1 - Processus générique de la Veille Stratégique VS

Au cœur du processus VASIC, se situe la « création collective de sens » (CCS). Une séance de travail collectif est organisée pour créer du « sens ajouté » et de la connaissance à partir d'informations qui jouent le rôle de stimuli. Les interactions entre les participants et les différentes mémoires (tacites et formelles) de l'entreprise vont aider à faire émerger du « sens ». Une séance de création collective de sens a pour résultat la formulation de conclusions temporaires (hypothèses plausibles) qui peuvent aboutir à des actions effectives [12] .

Pour aider les entreprises à créer collectivement du sens, l'équipe de veille stratégique du CERAG a proposé une méthode nommée « Puzzle » [13] . L'idée à l'origine est la métaphore du jeu de puzzle. Lors de la séance de création collective de sens, l'animateur apporte 5 ou 6 brèves (« une information ramenée à ses mots essentiels de façon à être très courte. Cette contrainte de taille résulte du fait qu'une brève est destinée à être projetée sur un écran » [15]). Ces brèves vont être utilisées comme des pièces d'un puzzle à construire. Contrairement au jeu de puzzle, on ne dispose pas d'un modèle ni de toutes les pièces. « La **méthode Puzzle**

consiste à construire collectivement un puzzle (ou plusieurs si nécessaire), sur l'écran de la salle de travail, en utilisant des informations à caractère anticipatif en guise de pièces, d'une part, et les suggestions et connaissances tacites que les participants explicitent alors, d'autre part » [15] .

La méthode Puzzle, appliquée des dizaines de fois dans divers organismes (entreprises et ministères), a toujours donné satisfaction. En revanche, ces mêmes interlocuteurs ont clairement indiqué que cette méthode ne pourrait pas être durablement utilisée dans leurs organismes : la préparation des informations nécessaires, dites informations brèves, demande trop de temps de travail humain. Alors que la recherche d'informations FULL *texts* (données brutes sous forme de textes) est désormais très rapide sur l'Internet, la préparation des brèves (c'est-à-dire le filtrage des données FULL *texts* pertinentes, la sélection des informations en relation avec le sujet à traiter, l'identification du passage bref utile pour produire une brève elle-même utile pour la création collective de sens), ne peut être faite que manuellement, du moins pour le moment, et nécessite trop de travail et donc des coûts importants causes de rejet de l'utilisation de la méthode. Le processus de veille est mis en échec à cause de la surcharge d'information qui entraine une impossibilité d'extraire de la connaissance [17] (surcharge cognitive) permettant d'anticiper une menace ou une opportunité d'affaires [11], [14] .

2.2 Présentation du cas « valorisation du CO_2 » et émergence de la problématique

La direction de l'entreprise appelée « *Durability* » (secteur de la chimie) a décidé d'explorer la possibilité d'exploiter le CO_2 en tant que matière première dans le but de diversifier ses activités dans une orientation stratégique d'avenir. Une séance de réflexion collective du comité de direction, a été décidée. L'ordre du jour tient dans la problématique suivante : *« Explorer la possibilité et la pertinence économique de valoriser le CO_2 en tant que matière première éventuelle ».*
La personne chargée de préparer les informations FULL *texts* susceptibles d'être utilisées au cours de la prochaine séance du comité de direction a recueilli 299 FULL *texts* pouvant correspondre à l'ordre du jour, mais il est hors de question que cette personne les lise tous dans le peu de temps dont elle dispose (surcharge de données). Cette personne sera nommée « animateur » par la suite.

Pour préparer la réunion, l'animateur doit effectuer les tâches suivantes :

- rechercher et extraire les FULL *texts* se rapprochant de l'ordre du jour et susceptibles de contenir de possibles informations à caractère anticipatif annonciatrices de changements dans l'environnement de l'entreprise,
- répondre au sujet du degré de fiabilité de chacun d'eux,
- faire de nombreux allers et retours d'un FULL *texts* à un autre pour accompagner les participants dans leurs réflexions et interactions.

Antérieurement de telles réunions de réflexion collective ont déjà été effectuées. Elles ont donné lieu à des échanges constructifs dont voici quelques exemples.

Exemples (verbatim) **–** Interactions entre participants, lors d'une séance de création collective de sens :

- « Le rapprochement de cette information-ci avec cette information-là me donne à penser que… ».
- « Ces deux informations voisines semblent *incohérentes*, à moins que… ».
- « Ce que tu es en train de dire *complète* bien ce qui est au tableau… C'est le chaînon manquant entre les deux informations... ».
- « Mais il y a une chose qui devrait nous donner à penser que… ».
- « En voyant ce que nous avons écrit au tableau *ça me fait penser que* la semaine dernière quelqu'un m'a dit que… ».
- « Il faudrait *compléter* en recherchant… ».
- « Sommes-nous certains que l'information que nous discutons est *fiable* ? »
- « Possédons-nous déjà des informations *'voisines'* de celle-ci, qui viendraient l'étayer ? ».

Le directeur général a décidé de rendre plus fréquentes de telles réunions de travail appelées « Création Collective de sens ou CCS » pour interpréter les informations à caractère anticipatif. Mais il exige que le **temps de préparation** et de manipulations des FULL *texts* soit **le plus bref possible** pour réduire les coûts administratifs, augmenter la réactivité et ne pas gêner les réflexions : sans un gain de temps significatif la veille stratégique sera abandonnée ! [9] .

Hypothèses : Parmi toutes les informations collectées, les cas suivants, tous ou seulement certains, pourraient se rencontrer :

1 - informations sans intérêt pour la problématique « proposée » par le management ;
2 - informations strictement redondantes (à tout point de vue) ;
3 – informations qui pourraient se compléter ;
4 – informations qui pourraient se contredire ;
5 – informations qui pourraient se conforter ;
6 – informations telles qu'il pourrait être envisagé des liens entre elles ;

Espoir : **SI** c'était possible de répartir en petits groupes les informations (FULL *texts*) trouvées préalablement (petits groupes pour lesquels on pourrait identifier les cas précédents) **ALORS** le grand nombre d'informations pourrait présenter une utilité, dès lors que les informations correspondant au cas 1 sont éliminées.

Hypothèse de recherche **: SI un logiciel permettait de rassembler rapidement/facilement les informations voisines d'une information donnée ALORS la veille anticipative stratégique utilisant des informations à caractère anticipatif gagnerait en efficience (en termes de coûts et de satisfaction des utilisateurs ainsi que de la hiérarchie).**

Ainsi la **problématique** se ramène à la question suivante : Peut-on concevoir un logiciel capable de rechercher, dans une surcharge de données, **les informations proches dites « informations voisines » afin de les répartir en petits groupes**?

Le concept de « voisinage » est présenté au paragraphe suivant.

N'ayant pas connaissance qu'un tel logiciel existe sur le marché [4] , la première étape de la recherche concerne la conception et la construction du logiciel ainsi que son test sur un premier cas expérimental. Cette étape est présentée dans le présent article.

2.3 Informations voisines

2.3.1 Définition :

Deux informations, au sein d'une base de données se rapportant au même sujet donné (ou ordre du jour), sont dites voisines si on peut les rapprocher selon les trois dimensions suivantes :
- les mots en commun et/ou
- les synonymes (par rapport à un certain dictionnaire de synonymes) et/ou
- les mots cooccurrents.

2.3.2 Utilité du concept de voisinage

Le concept « Informations voisines » est utile lorsqu'il s'agit : de rapprocher deux informations (ou plus) à caractères potentiellement anticipatif, pas nécessairement écrites avec les mêmes mots, dans le but de :
- fiabiliser l'une d'elles (ou plusieurs),
- compléter l'une d'elles (ou plusieurs),
- mettre en évidence une incohérence ou une contradiction entre deux informations,
- faciliter l'interprétation de l'ensemble de ces informations,
- mettre en évidence un prolongement de la problématique que l'on ignorait,
- d'envisager des liens entre des informations (complémentarité, incohérence, contradiction, etc.).

2.3.3 Qui est concerné par la recherche d'informations voisines ?

Est concernée principalement l'animateur en charge de la gestion de la base de données où sont stockés les FULL *texts* résultant de l'interrogation des diverses sources surveillées. La hiérarchie aura indiqué à l'animateur la *question (ordre du jour)* abordée lors de la prochaine séance de travail collectif en vue d'éclairer une éventuelle prise de décision stratégique. La base de données est supposée contenir de potentielles informations à caractère anticipatif. Sachant que les bases de données peuvent être énormes et que la moindre recherche sur l'Internet peut produire plusieurs centaines de FULL *texts* on comprend le stress que fait peser, sur l'animateur, la *pression de temps* exercée par la hiérarchie ainsi que l'anxiété qui en résulte [3] .

3 Prototype Alhena

3.1 Mesure de voisinage

Le voisinage n'est pas binaire. Le voisinage est une grandeur **mesurable** (voir plus bas). Elle varie entre deux extrêmes : « pas de voisinage du tout » à « identiques ». Afin de mesurer le voisinage de textes et de proposer à l'animateur une visualisation des résultats, nous procédons en trois étapes :
- représentation informatique des textes,
- calcul de la mesure de voisinage,
- construction graphique des résultats.

3.1.1 Représentation informatique des textes

Dans cette première étape, nous transformons les textes de manière à pouvoir effectuer des calculs. Cette transformation est réalisée en 2 temps :
- dans un premier temps, nous « lemmatisons » les textes à l'aide du lemmatiseur **TreeTagger** : cet outil permet de mettre les verbes à l'infinitif, les noms et adjectifs au masculin singulier ; si un mot n'existe pas ou s'il est mal orthographié ou encore s'il s'agit d'un nom propre alors il est inchangé. Nous établissons ainsi une liste de mots lemmatisés pour chaque texte.
- dans un deuxième temps, nous supprimons les mots-outils à l'aide d'un dictionnaire de mots-outils (on parle d'antidictionnaire). On entend par mot-outil les articles, les auxiliaires (être et avoir), les chiffres, les symboles (comme $ ou €), certains adverbes, … Notre antidictionnaire n'est pas figé et il peut être enrichi à tout moment.

3.1.2 Calcul de la mesure de voisinage

Pour comparer nos textes, nous avons construit une mesure qui estime le voisinage entre deux textes. Les calculs des mesures entre textes utilisent les listes de mots lemmatisés de chaque texte. La mesure entre deux textes s'appuie sur trois critères :

- les mots en commun : un même mot est présent dans chacune des listes de mots lemmatisés des deux textes,
- les synonymes : un mot d'une des listes de mots lemmatisés a un synonyme dans l'autre liste,
- la cooccurrence : un mot d'une des listes de mots lemmatisés apparaît très fréquemment dans l'ensemble des textes de la base de données avec un mot de l'autre liste.

Description du calcul de la mesure[1] : prenons deux textes que nous notons T_i et T_j. Nous établissons la mesure de voisinage entre T_i et T_j de la manière suivante :

- nous établissons les listes L_{Ti} et L_{Tj} des mots lemmatisés, inconnus ou mal orthographiés de respectivement T_i et T_j,
- pour chaque mot de L_{Ti}, nous ajoutons :
  - 0 si le mot est aussi présent dans L_{Tj},
  - sinon *valeur_synonyme*[2] si le mot a un synonyme dans L_{Tj},
  - sinon *valeur_cooccurrence_min*[1] : on cherche le mot dans L_{Tj} pour lequel la *valeur_cooccurrence*[1] est la plus petite,
- nous inversons les rôles de L_{Ti} et L_{Tj} et nous procédons à nouveau comme ci-dessus,
- pour terminer, nous ajoutons la somme sur les mots de L_{Ti} avec la somme sur les mots de L_{Tj} et nous divisons par 2.

Nous avons calculé la mesure de voisinage entre chaque texte : nous avons ainsi construit une matrice carrée de taille nombre textes. La ligne i colonne j donne la mesure de voisinage entre le texte T_i et le texte T_j. Cette matrice contient beaucoup d'informations mais elle est difficilement exploitable. Comme notre mesure est une mesure de proximité (i.e. plus la mesure est petite, plus les textes sont proches), nous avons cherché le(s) minimum(s) sur chacune des lignes i.e. pour chacun des textes, nous avons cherché son plus proche voisin. Cette recherche du (des) minimum(s) pour un texte est réalisée, bien sûr, sans tenir compte de la mesure d'un texte à lui-même (qui par définition est nulle et qui se trouve sur la diagonale de la matrice des mesures de voisinage). Nous construisons un graphe à partir de cette recherche : nous traçons une flèche entre un texte et le(s) texte(s) qui réalise(nt) la mesure minimale (voir Figure 2). Un exemple de graphique obtenu sur 11 textes est donné à la Figure 3.

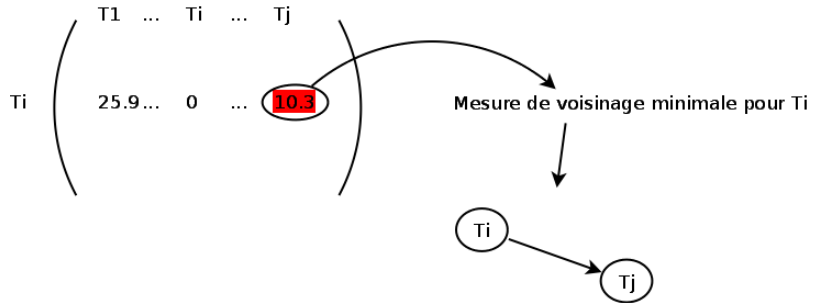


Figure 2- Réalisation de la mesure minimale et construction du graphe

[1] La description formelle de la mesure de voisinage est donnée en annexe.
[2] La valeur_synonyme, la valeur cooccurrence sont basées sur un calcul de la distance Google établit par Cilibrasi et Vitanyi [5] .

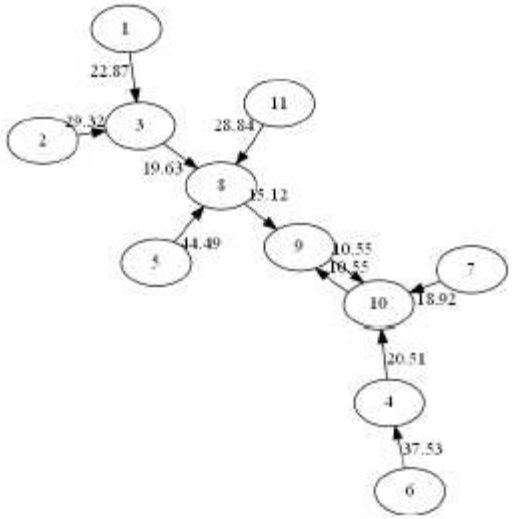


Figure 3- Exemple de graphe obtenu avec la mesure de voisinage

Le graphe ainsi construit présente les propriétés suivantes :

- en partant des extrémités du graphe et en suivant les flèches jusqu'au « nucléus » (« double flèches » ou cycle ; par exemple les textes 9 et 10 de la Figure 3), le mesure de voisinage est décroissante
- le nucléus réalise la mesure minimale
- les textes situés aux extrémités du graphe comptent plus de mots que les textes du nucléus : en suivant les flèches dans le graphe, les textes sont de plus en plus courts.

3.2 Prototype Alhena

Le prototype Alhena a été créé pour « opérationnaliser » la mesure de voisinage et en faire ainsi une connaissance « actionnable » [1] . Un des objectifs était de proposer un outil d'aide à la lecture des FULL *texts* en proposant des regroupements de FULL *texts* voisins. Nous avons décidé de nommer notre prototype Alhena en référence à une étoile de la constellation des Gémeaux car les graphes obtenus suggèrent un ciel étoilé où l'on distingue des constellations. Dans un premier temps, les textes sont insérés dans la base puis un algorithme de calcul de la mesure de voisinage et de traçage du graphe est lancé. Pour obtenir les graphiques nous nous appuyons sur le logiciel **GraphViz**. Le graphe obtenu est appelé galaxie et les composantes de ce graphe sont nommées constellations locales (CL).

3.3 Cas « valorisation du CO2 »

Une séance de réflexion collective, réunissant divers directeurs (comité de direction), a été décidée. L'ordre du jour est libellé ainsi : « Explorer la possibilité et la pertinence économique de valoriser le CO2 en tant que matière première éventuelle pouvant être valorisée ». L'animateur dispose 299 FULL *texts* pouvant correspondre à l'ordre du jour, mais il est hors de question que l'animateur les lise toutes dans le peu de temps dont il dispose (surcharge d'information) : l'animateur doit donc extraire les plus pertinents et le nombre de ceux-ci ne doit pas dépasser la quinzaine afin de pouvoir être exploités lors de la réunion.

Pour la réunion, l'animateur doit donc se mettre en condition :

- de présenter les FULL *texts* sélectionnés de façon aussi visuelle que possible,
- de répondre rapidement aux demandes que pourraient formuler, « au fil de l'eau », les participants,
- d'accompagner le déroulement des interactions, sans briser le rythme de celles-ci, entre les participants en projetant les FULL *texts* éventuellement susceptibles d'aider à la réflexion collective,
- de répondre, au fur et à mesure et rapidement, à d'éventuelles questions du genre : « *Cette information est-elle fiable ? Disposons-nous d'une information qui viendrait la compléter ? Disposons-nous d'une information qui viendrait contredire ou infirmer … ?* ».

C'est dans le but d'apporter une aide efficiente, pour répondre à de telles conditions, qu'a été conçu et construit le prototype, nommé Alhena présenté si après. Il a fait l'objet d'une première expérimentation qui sert de support au cas présenté ci-après.

Au cours de la préparation de la future séance de travail, prévue pour le lendemain (pression du temps), la première tâche de l'animateur est de découvrir le contenu des 299 FULL *texts*. Il dispose de très peu de temps pour cela. Il utilise donc Alhena. Voici la suite des opérations qu'il a effectuées.

3.3.1 Résultat visuel

Alhena affiche la représentation visuelle illustrée par la Figure 4 en forme de « **galaxie** » globale dans laquelle les 299 FULL *texts* sont identifiables par leur numéro. L'animateur observe trois types de formes graphiques : des doubles flèches ; des petites **constellations locales centrées** sur leur nucléus (les doubles flèches) (l'une d'elles sera commentée plus bas) ; des bras des constellations locales, bras constitués des suites de FULL *texts* reliés entre eux par une simple flèche.

Grâce aux propriétés de la mesure de voisinage et des graphes obtenus, il peut suffire de lire seulement le texte du nucléus :

- Soit pour décider d'abandonner la lecture de tous les FULL *texts* de la constellation (gain de temps, réduction de la surcharge)
- Soit pour être alerté sur l'utilité de lire au moins les FULL *texts* de la première couronne du nucléus : augmentation de l'attention.

La démarche effectuée par l'animateur a été la suivante. Pour commencer, l'animateur clique sur une constellation locale que nous noterons CL (par exemple celle entourée d'un trait dans la Figure 4). Il obtient ainsi une page avec la représentation de la constellation locale, un tableau avec le nuage de mots de CL ainsi qu'un tableau contenant les nuages de mots pour chaque bras de la constellation. Le nuage de mots de CL permet à l'animateur d'avoir une idée générale sur le sujet abordé par CL. Les nuages de mots des branches donnent à l'animateur un aperçu de la manière dont le sujet de CL est abordé par chaque branche (voir Figure 5).

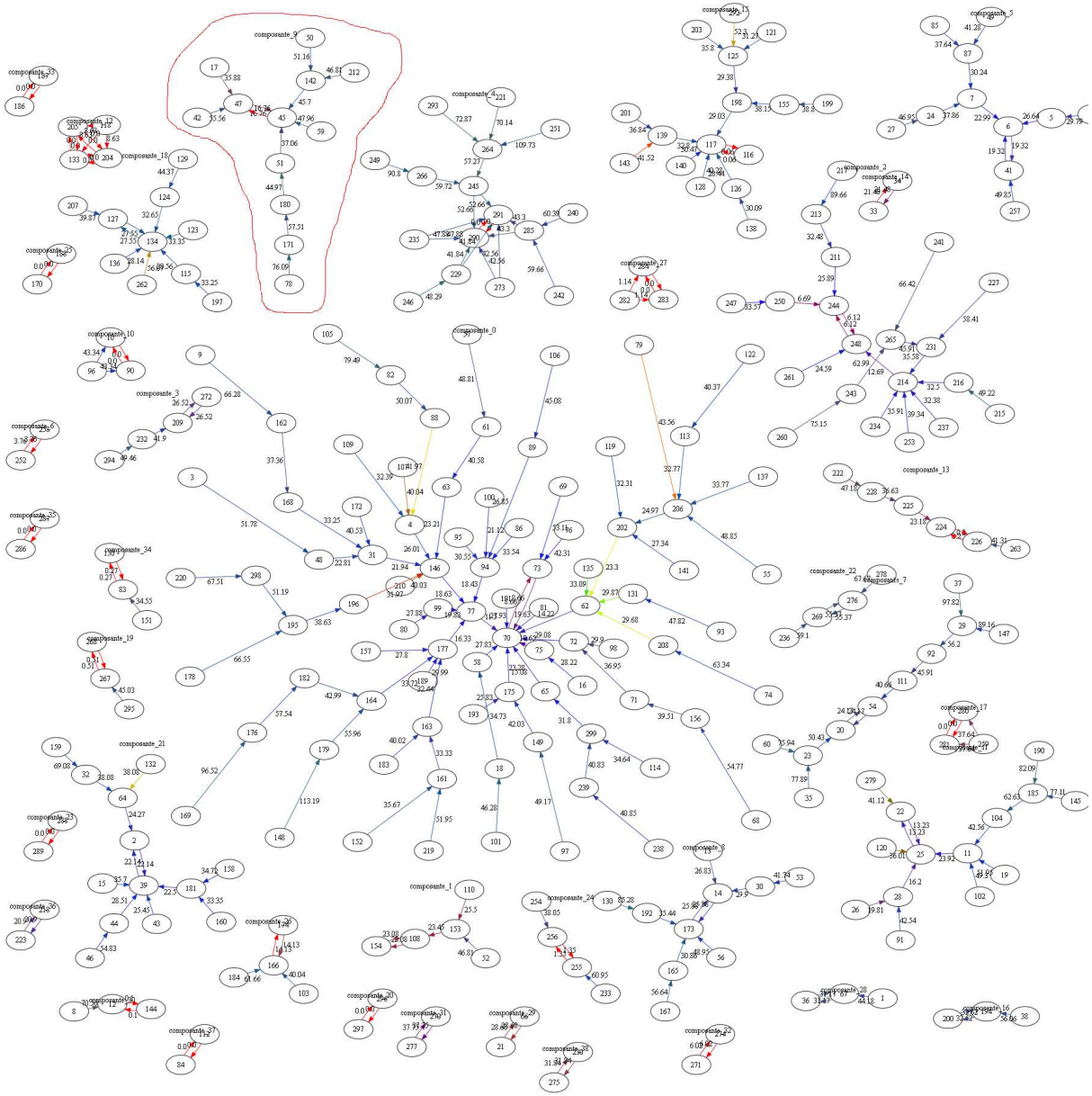
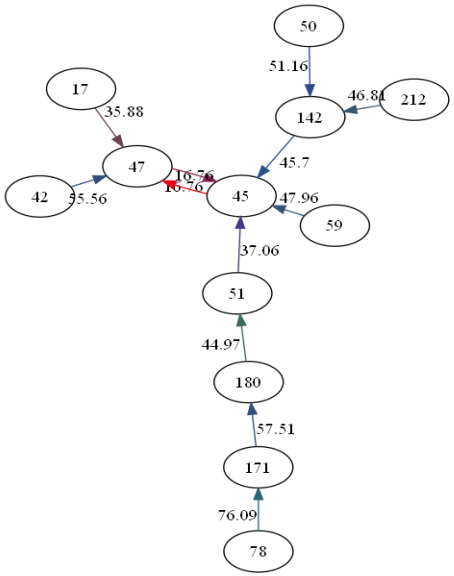


Figure 4 - Galaxie globale proposée par Alhena



biocarburants biochimique biodiésel budget captage carburant cecr cellulosique charbon charge chef créer devoir dollar déchet enerkem estrie etat europe européen financement gaz greencentre greenfield greenline génération inc. industries innovation kingston leaf maroc norme papetier pays production produire produit recherche science stockage synthèse tourner tunis tunisie tunisien université usine étape éthanol

47,142,212,45,	accélérer agricole américain an avancer base biocarburants canada captage carburant cellulosique charbon chef co2 commercial concerner créer devoir diminuer département dépendance emploi enerkem exploitant gallon ges gouvernement greenfield groupe incitation interministériel litre montréal norme obama ontario production produire propos propre rapport renouvelable réduire société stockage stratégie technologie usine énergie éthanol
47,51,180,78,171,45,	arriver avancer biocarburants budget canada cecr cellulosique charge chef chimique chômage créer devoir dollar enerkem entreprise estrie etat europe européen exemple expertise exportation financement greencentre greenfield innovantes innovation jeune kingston maroc mondial pays politique production public pôle rce recherche relance science technologique tourner tunis tunisie tunisien université usine éthanol étranger
47,45,17,	biocarburants biochimique carburant cellulosique chef chimique commercial conditionné devenir déchet démarrage enerkem exploiter file gaz ges greenfield génération matière montréal ontario production produire produit québec renouvelable réduire situer société synthèse technologique transformer usine vert voiture westbury étape éthanol évoluer îlot
47,50,142,45,	américain an biocarburants bois canada cellulosique chef commercial concerner dollar déjà département enerkem euro fabriquer ges grand greenfield montréal ontario papetier papier pouvoir production propos rapport recherche réduire sembler société stora suède technologie usine énergie éthanol
47,59,45,	accepter actualité agir air alimentation anciens and animé biocarburants business canada canadien capacité carburant catalyse cellulosique charbon chef classer chimie commercial compagnies compter construction consentir effectuer enerkem entreprise faillois ges greenfield installation montréal ontario organique partir poteau production propos réacteur sherbrooke société technologie technologies téléphone urbain usine équipement éthanol

Figure 5- Informations sur la constellation locale présentée

3.3.2 Nucléus

L'animateur pointe d'abord l'un des nucléus (doubles flèches).
Prenons l'exemple de la double flèche reliant les FULL *texts* : 45 <--> 47, au centre de la constellation entourée par un trait rouge dans la Figure 4.
Il observe que chacune des flèches porte un nombre. Ce nombre est le même, soit 16,76, pour les deux flèches. Il est le plus faible de la constellation locale, par conséquent 45 et 47 sont les plus voisins au sens défini plus haut. En effet, c'est une propriété générale, le nucléus réalise toujours la plus petite mesure de chaque constellation locale.
L'animateur clique sur 45, instantanément le FULL *text* 45 complet s'affiche sur la moitié gauche de l'écran, tandis sur la partie droite s'affiche le texte complet 47.
L'animateur peut ainsi comparer les textes 45 et 47 s'il en est besoin (voir Figure 6).

La comparaison est facilitée car les mots communs aux textes 45 et 47 sont surlignés en couleur saumon par Alhena. S'il y a des synonymes ils sont affichés en bleu, tandis que d'éventuelles cooccurrences entre des mots sont indiquées par soulignage. De plus le logiciel indique, dans la marge horizontale en haut de l'écran, les mots les plus fréquents dans les deux textes. Ici il s'agit des mots : Enerkem, GreenField (tous deux industriels de la chimie), éthanol et cellulosique.

Plusieurs conclusions émergent des constats que peuvent faire l'animateur et les participants :

- 45 et 47 ont un grand nombre de mots en commun mais ils ne sont pas des doublons car ils ne sont pas constitués du même **nombre de mots** (381 pour FULL *text* 45 et 835 pour FULL *text* 47) ; leurs **sources d'émission** ne sont pas les mêmes, pas plus que leurs **dates d'émission** (11 mars 2008 pour 47 et 18 mars pour 45). L'animateur peut conclure que 45 et 47 se fiabilisent mutuellement, ou tout au moins se confortent de façon significative. L'animateur sera ainsi moins démuni si la question de la fiabilité lui est posée en cours de la séance de travail collectif.

- deux noms d'acteurs apparaissent : *Enerkem et GreenField*, dont personne n'avait parlé jusqu'ici chez Durability. S'agirait-il de concurrents dont on n'avait pas pris conscience, concernant la piste stratégique CO_2 explorée ?

- Durability explore la possibilité de valoriser le CO_2 en produisant de l'éthanol au moyen d'algues. Mais ici apparaît l'éthanol cellulosique : s'agirait-il d'une piste concurrente… qui ferait appel à une technologie concurrente ? Faut-il s'en inquiéter ? L'attention est alertée sur des interrogations que les dirigeants n'avaient peut être pas encore envisagées.

- La comparaison de 45 et 47 est largement facilitée puisque le logiciel surligne en couleur saumon tous les mots identiques dans 45, à gauche de l'écran et dans 47, à droite de l'écran.

En résumé,
- le logiciel attire l'attention sur un point d'entrée au sein des 299 FULL *texts* : le nucléus 45/47,
- en pointant sur le nucléus, le logiciel suggère de nombreux éléments pour alimenter la réflexion stratégique du comité de direction de Durability,
- le temps nécessaire est très court comparé à ce qu'il serait si l'animateur avait dû tout faire manuellement, ce qui constitue un argument décisif aux yeux de la direction. Mais ce n'est pas le seul apport du prototype Alhena.

Retour vers la constellation

45	47
biocarburants canada cellulosique chef commercial enerkem greenfield production société technologie usine éthanol	canada cellulosique enerkem greenfield production usine éthanol
ÉTHANOL GREENFIELD ET ENERKEM ANNONCENT UN PROJET COMMUN DE PRODUCTION D'ÉTHANOL CELLULOSIQUE 381 mots 18 mars 2008 eSource Canada Business News Network Anglais Copyright 2008 Business Information Group. All Rights Reserved. Éthanol GreenField et Enerkem viennent de signer une entente de principe visant la production d'éthanol cellulosique sur une échelle commerciale. « Nous sommes ravis de travailler avec Enerkem pour faire de l'éthanol cellulosique une réalité commerciale au Canada », a déclaré Bob Gallant, président et chef de la direction d'Éthanol GreenField. « Les consommateurs canadiens veulent une solution de rechange plus verte et peu coûteuse aux combustibles fossiles et GreenField répond à cette demande en accroissant de façon significative ses activités de production de nouveaux biocarburants», a dit Frank Dottori, directeur général de la division d'éthanol cellulosique de GreenField. Selon l'entente, les deux sociétés collaboreront, à parts égales, à des projets répartis dans des régions spécifiques choisies et visant la conception, la construction et l'exploitation d'usines de fabrication d'éthanol cellulosique utilisant la technologie d'Enerkem. L'emplacement de la première usine a été choisi au Canada et sera annoncé au cours des prochaines semaines. Une deuxième usine est en cours de développement. La technologie d'Enerkem transforme la biomasse comme les déchets solides municipaux triés et les résidus forestiers en éthanol cellulosique et autres biocarburants. Elle permet l'élimination de plus de deux tonnes de gaz à effet de serre par tonne de matières résiduelles utilisées dans le procédé. Les fondateurs de la société ont consacré plusieurs années à développer cette technologie de gazéification. L'usine pilote d'Enerkem, qui, depuis 2003, cumule plus de 3 000 heures d'opération, fabrique du gaz de synthèse, du méthanol et de l'éthanol cellulosique. De plus, la société construit actuellement une usine de démonstration de taille commerciale pour la production d'éthanol cellulosique à Westbury, au Québec. « Ce partenariat est une étape déterminante dans l'atteinte de l'objectif d'Enerkem qu'est la commercialisation d'éthanol cellulosique », a dit M. Vincent Chornet, président et chef de la direction d'Enerkem. « En unissant nos forces à celles d'Éthanol GreenField, nous sommes confiants de devenir des chefs de file canadiens dans la production et la distribution de biocarburants de nouvelle génération. L'expérience de GreenField dans la construction et l'exploitation d'usines industrielles sera essentielle à l'expansion de notre production », conclut M. Chornet.	Éthanol GreenField et Enerkem annoncent un projet de production d'éthanol cellulosique d'envergure commerciale 835 mots 11 mars 2008 14:44 Canada Newswire Français Copyright © 2008 Canada NewsWire Ltd, tous droits réservés. TORONTO and MONTRÉAL, le 11 mars /CNW/ -- TORONTO and MONTRÉAL, le 11 mars /CNW/ - Éthanol GreenField, le plus gros producteur d'éthanol au Canada, et Enerkem, une entreprise de premier plan dans le domaine des technologies de gazéification et de catalyse, ont signé une entente de principe visant la production d'éthanol cellulosique sur une échelle commerciale. "Nous sommes ravis de travailler avec Enerkem pour faire de l'éthanol cellulosique une réalité commerciale au Canada", a dit Bob Gallant, président et chef de la direction d'Éthanol GreenField. "Les consommateurs canadiens veulent une solution de rechange plus verte et peu coûteuse aux combustibles fossiles et GreenField répond à cette demande en accroissant de façon significative ses activités de production de nouveaux biocarburants", a dit Frank Dottori, directeur général de la division d'éthanol cellulosique de GreenField. Selon l'entente, les deux sociétés collaboreront, à parts égales, à des projets répartis dans des régions spécifiques choisies et visant la conception, la construction et l'exploitation d'usines de fabrication d'éthanol cellulosique utilisant la technologie d'Enerkem. L'emplacement de la première usine a été choisi au Canada et sera annoncé au cours des prochaines semaines. Une deuxième usine est en cours de développement. La technologie d'Enerkem transforme la biomasse comme les déchets solides municipaux triés et les résidus forestiers en éthanol cellulosique et autres biocarburants. Elle permet l'élimination de plus de deux tonnes de gaz à effet de serre par tonne de matières résiduelles utilisées dans le procédé. Les fondateurs de la société ont consacré plusieurs années à développer cette technologie de gazéification. L'usine pilote d'Enerkem, qui, depuis 2003, cumule plus de 3 000 heures d'opération, fabrique du gaz de synthèse, du méthanol et de l'éthanol cellulosique. De plus, la société construit actuellement une usine de démonstration de taille commerciale pour la production d'éthanol cellulosique à Westbury, au Québec. "Ce partenariat est une étape déterminante dans l'atteinte de l'objectif d'Enerkem qu'est la commercialisation d'éthanol cellulosique", a dit Vincent Chornet, président et chef de la direction d'Enerkem. "En unissant nos forces à celles d'Éthanol GreenField, nous sommes confiants de devenir des chefs de file canadiens dans la production et la distribution de biocarburants de nouvelle génération. L'expérience de GreenField dans la construction et l'exploitation d'usines industrielles sera essentielle à l'expansion de notre production", conclut M. Chornet. À propos de l'éthanol --- L'éthanol est un carburant renouvelable fabriqué à partir de céréales comme le maïs et le blé, ou à partir de la cellulose qu'on trouve dans les plantes et la biomasse. L'éthanol est peu coûteux et offre des avantages environnementaux uniques. Le modèle GH Genius de Ressources naturelles Canada indique que l'éthanol produit à partir du maïs permet de réduire les émissions de gaz à effet de serre (GES) de 40 à 60 pour cent comparativement à l'essence. L'éthanol cellulosique peut réduire les GES de 87 pour cent selon le modèle GREET du ministère de l'Énergie des États-Unis. L'engagement du gouvernement fédéral à ce que l'essence comprenne une moyenne de cinq pour cent d'éthanol d'ici 2010 permettra de réduire les émissions de GES d'une quantité équivalant au retrait d'un million de voitures des routes canadiennes chaque année. À propos d'Enerkem --- Enerkem, dont le siège social est situé à Montréal, a des bureaux d'ingénierie à Sherbrooke. C'est un chef de file dans le développement de biocarburants cellulosiques. La technologie de gazéification, de conditionnement du gaz synthétique et de catalyse d'Enerkem transforme les déchets solides municipaux triés et les résidus forestiers en éthanol cellulosique et autres biocarburants. La société exploite une usine pilote depuis 2003 et construit actuellement au Canada une usine de démonstration de taille commerciale pour la production d'éthanol cellulosique. www.enerkem.com À propos d'Éthanol GreenField --- Éthanol GreenField, auparavant Les Alcools de commerce, est le principal producteur d'éthanol du Canada. Chaque année, la société produit 250 millions de litres d'éthanol à ses usines de Chatham et Tiverton, en Ontario, et de Varennes, au Québec. Sa plus grosse usine jusqu'à présent, d'une capacité de 200 millions de litres, à Johnstown, en Ontario, sera opérationnelle en décembre 2008. Une usine de 145 millions de litres est en cours de développement à Hensall, en Ontario. GreenField participe activement au processus de développement d'une technologie biochimique visant à produire de l'éthanol cellulosique à ses installations de pointe situées à Chatham, en Ontario. Le carburant d'Éthanol Greenfield est offert dans plus de 1 500 stations service, partout au Canada. Pour plus d'information, veuillez visiter le www.greenfieldethanol.com . Melissa Armstrong, Éthanol GreenField, (416) 304-1700, poste 8431, m.armstrong@greenfieldethanol.com; Marie-Hélène Labrie, Enerkem Inc., (514) 875-0284, poste 224, mlabrie@enerkem.com

mot trouvé dans les 2 textes
synonyme trouvé
mots rapprochés par la distance google

Figure 6- Visualisation et comparaison entre les textes 45 et 47

3.3.3 Petite couronne autour du nucléus

L'animateur se demande si les informations 45 /47 peuvent être complétées, voire fiabilisées davantage. Il examine alors les FULL *texts* qui constituent la couronne rapprochée autour de 45/47 dans la constellation affichée sur l'écran, soit 51, 142, 17, 42 et 59. En cliquant tour à tour sur chacun de ces numéros le texte complet du FULL text apparaît, de même que les mots les plus fréquents, dans la marge horizontale en haut de l'écran.
L'animateur peut constater, en très peu de temps, que tous les textes de la couronne de 45 / 47 (sauf 42) apportent des éléments qui viennent compléter les renseignements concernant les acteurs Enerkem et GreenField, d'une part, et l'éthanol cellulosique, d'autre part. Ces éléments pourront donc contribuer à rassurer les membres du comité de direction au sujet de la fiabilité des informations 45 /47. Ils pourront aussi alerter leur vigilance sur les concurrents potentiels Enerkem et Greenfield.

3.3.4 Bras de constellation

Les participants pourront demander à l'animateur pourquoi des FULL *texts* sont reliés entre eux pour constituer un bras de constellation locale. Par exemple 78, 171, 180, 51 constituent le bras le plus long de la constellation locale 45/47. Si l'animateur clique sur 78, Alhena affiche sur l'écran les textes complets des FULL *texts* 78 et de 171. Il indique également : les mots communs aux deux textes (couleur saumon) ; les synonymes et les éventuelles cooccurrences (ainsi que déjà mentionné plus haut). Puis l'animateur peut cliquer sur 171, par exemple : apparaissent alors les textes de 171 et de 180, et ainsi de suite jusqu'à arriver au nucléus.

N.B. Les textes par lesquels on transite depuis l'extrémité d'un bras de constellation locale jusque vers le nucléus de celle-ci sont de plus en plus courts et les mots principaux occupent une place relative de plus en plus importante : dans le cas présent ce sont les mots Enerkem et éthanol cellulosique.

3.4 Synthèse

Pour les explications précédentes, nous avons choisi d'utiliser les informations 45/47, reliées par une double flèche et situées au centre d'une constellation locale (Figure 5). Nous avons pu montrer ainsi que l'existence d'une telle constellation locale est de nature à attirer l'attention sur une sous thématique qui n'avait peut-être jamais évoquée par les dirigeants de Durability jusque-là, mais dont la découverte peut susciter une réaction stratégique. Le « champ de vision » de la hiérarchie est augmenté là où il y avait un angle mort ! Mais la constellation globale contient d'autres galaxies locales qui doivent être examinées de la même façon.

L'examen de la totalité des nucléus de la galaxie (Figure 4) a été fait par l'animateur au cours de sa préparation. Le temps nécessaire a été d'environ deux heures : une heure de fonctionnement du logiciel (sans intervention humaine) puis une heure de « travail humain » pour l'analyse de la galaxie de la part de l'animateur. Le résultat est que seul un autre nucléus s'est également montré possiblement intéressant : le nucléus 70/73, il attire l'attention vers la piste « Industrie Alimentaire » pour la valorisation le CO2. Par la suite cette piste a déclenché une surprise parmi le comité de direction : il ne l'avait jamais évoquée jusqu'à ce moment. Les questions stratégiques suivantes ont été soulevées : « Faut-il s'intéresser à cette piste ? Faudrait-il envisager des partenariats ? Quels pourraient être les débouchés dans la décennie à venir ? Etc. ».

4 Conclusion

L'expérimentation effectuée montre que l'hypothèse de recherche présentée plus haut est en bonne voie de **validation** : elle encourage à répliquer la démarche sur de nouveaux terrains d'application.

4.1 Avantages attendus du prototype

L'objectif visé était d'apporter une aide pour surmonter la paralysie occasionnée par les gros volumes de données numériques et, par voie de conséquence, pour diminuer les coûts administratifs de la veille anticipative stratégique ainsi que la surcharge cognitive. L'expérimentation effectuée permet de constater :

- Gain de temps considérable pour l'exploration des FULL *texts* tirés des sources Internet (réduction d'une vingtaine d'heures à deux heures environ, dans le cas de l'expérimentation effectuée) ;
- Une facilitation pour rechercher des informations de nature à se fiabiliser, ou se conforter, mutuellement ;
- Une facilitation pour suggérer des liens entre des informations (construction des puzzles utilisant les 'pièces' d'information que sont les signaux faibles) de la part des participants au groupe de travail chargé d'interpréter les signaux faibles (lien d'incohérence entre deux informations ; lien de complémentarité ; lien de causalité entre deux informations, etc.)
- le prototype ne nécessite pas de fichiers textes structurés (par exemple avec des balises pour reconnaître le titre, l'auteur, la date, …). Cette spécificité nous permet de comparer des textes issus de sources différentes.
- Alhena est applicable pour n'importe quelle langue si l'on dispose d'un lemmatiseur pour celle-ci.

4.2 Limites actuelles de l'instrumentation du concept

- Alhena est, en l'état actuel du prototype, incapable de traiter des masses colossales de textes (par exemple le web) dans un temps raisonnable : pour le cas CO2, il propose une représentation graphique sur 300 textes au bout d'une heure. Mais le code pourra être optimisé pour traiter, dans un avenir proche, plus de données,
- Le prototype se heurte au problème d'encodage des fichiers : Alhena nécessite des fichiers txt au format accepté par le lemmatiseur,
- Alhena est un outil d'aide à la lecture pour l'animateur : il propose à un moment donné une classification des textes parmi toutes celles possibles. La classification proposée n'est sans doute pas la « proposition idéale »,
- Le prototype ne propose pas d'effectuer des recherches basées sur la sémantique ou la syntaxe : néanmoins Alhena ne s'attachant pas à la syntaxe ou à l'orthographe regroupe même des textes qui comportent des erreurs de syntaxe ou qui sont sémantiquement mauvais,
- Alhena n'est pour l'instant pas un outil multilingue : il travaille sur une seule langue à la fois.
- Nous n'avons réalisé qu'un trop petit nombre d'expérimentations pour le moment.

Dans le futur, Alhena pourra être amélioré en ajoutant le multilinguisme. De même, un module syntaxique et sémantique pourrait aider l'animateur à reconnaître des informations contradictoires. Enfin, nous envisageons de tester notre hypothèse de recherche en répliquant les expérimentations sur de nouveaux terrains d'application.

5 Bibliographie

[1] ARGYRIS C. *Actionable Knowledge: Design Causality in the Service of Consequential Theory*. Journal of Applied Behavioral Science. 32(4) , 1996, p390‑406.

[2] BAWDEN D. *Information overload*. Library & information briefings. (92) p1‑15.

[3] BAWDEN D, ROBINSON L. *The dark side of information: overload, anxiety and other paradoxes and pathologies*. Journal of Information Science. 35(2) , 4 janv 2009, p180‑191.

[4] BONDU J. *Benchmark des plateformes de veille : Choisir son outil [Internet]*. 2010. Disponible sur: http://www.inter-ligere.com/article-benchmark-des-plateformes-de-veille-choisir-son-outil-54266992.html

[5] CILIBRASI RL, VITANYI PMB. *The Google Similarity Distance*. IEEE Transactions on Knowledge and Data Engineering. 19(3) , 2007, p370‑383.

[6] COUDOL D, GROS S. *AgentIntelligent.com - Veille Strategique - Définition et Objectifs [Internet]*. [cité 24 avr 2013]. Disponible sur: http://www.agentintelligent.com/veille/veille_strategique.html

[7] DENIS J, ASSADI H. *Les usages de l'e-mail en entreprise. Efficacité dans le travail ou surcharge informationnelle ?* Le travail avec les technologies de l'information. , 2005, p135‑155.

[8] GANTZ J, REINSEL D. *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East [Internet]*. IDC - EMC; déc 2012. Disponible sur: http://www.emc.com/leadership/digital-universe/iview/index.htm

[9] HEMP P. *Death by information overload*. Harvard business review. 87(9) , sept 2009, p82‑89, 121.

[10] ISAAC H, CAMPOY É, KALIKA M. *Surcharge informationnelle, urgence et TIC. l'effet temporel des technologies de l'information*. Management & Avenir. n° 13(3) , 1 juin 2007, p149‑168.

[11] LAFAYE C, BERGER-DOUCE S. *Veille stratégique en petite entreprise : proposition de la notion d'intelligence collective entrepreneuriale*. Revue de l'Entrepreneuriat. Vol. 11(2) , 28 févr 2013, p11‑30.

[12] LESCA H. *Veille stratégique : La méthode L.E.SCAnning*. Management et Société (EMS); 2003.

[13] LESCA H. *Le problème crucial de la veille stratégique : la construction du «puzzle»*. Réalités industrielles. (AVR) p67‑71.

[14] LESCA H, KRIAA-MEDHAFFER S, CASAGRANDE A. *La surcharge d'information causée par l'Internet : Un facteur d'échec paradoxal largement avéré : Veille stratégique- Cas concrets, retours d'expérience et piste de solutions*. La Revue des Sciences de Gestion. 5(6) , 2010, p245‑246.

[15] LESCA H, LESCA N. *Les signaux faibles et la veille anticipative pour les décideurs : Méthodes et applications*. Hermes Science Publications; 2011.

[16] MLAIKI A, KALIKA M, KEFI H. *Facebook …encore, encore ! Rôle de l'affect, de l'habitude et de la surcharge informationnelle dans la continuité d'utilisation des réseaux sociaux numériques*. , 2011 [cité 24 mai 2013],. Disponible sur: http://basepub.dauphine.fr/xmlui/handle/123456789/7963

[17] NELSON MR. *We have the information you want, but getting it will cost you!: held hostage by information overload.* Crossroads. 1(1) , sept 1994, p11‑15.

[18] SAUVAJOL-RIALLAND C. *La surcharge informationnelle dans l'organisation : les cadres au bord de la « crise de nerf »*. Le magazine de la communication de crise et sensible (Observatoire International des Crises) [Internet]. 19 , 2010 [cité 24 avr 2013],. Disponible sur: http://www.communication-sensible.com/articles/article229.php

6 Annexe :

Définition de la mesure de voisinage basée sur la distance définie par Cilibrasi et Vitanyi [5]

Calcul de la distance de Cilibrasi et Vitanyi

Définition

Cilibrasi et Vitanyi établissent une distance entre deux termes M_l et M_k basée sur le nombre de pages trouvées/indexées par Google contenant M_l, puis contenant M_k, contenant M_l et M_k ainsi que le nombre total de pages indexées par Google. Leur distance, notée DCV, est définie par :

$$DCV(M_l, M_k) = \frac{\max(\log(f(M_l), \log(f(M_k))) - \log(f(M_l, M_k))}{\log(M) - \min(\log(f(M_l), \log(f(M_k))))}$$

avec :

- $f(x)$ = nombre de textes de la base contenant x
- $f(x, y)$ = nombre de la base contenant x et y
- M = nombre de textes dans la base.

Définition de la mesure de voisinage

Soit

- C un corpus,
- T_i et T_j deux textes
- L_{T_i} et L_{T_j} les listes des mots lemmatisés, inconnus ou mal orthographiés de respectivement T_i et T_j
- On note M_l un mot lemmatisé ou inconnu ou mal orthographié.

La mesure de voisinage est définie par :

$$MV(T_i, T_j) = \frac{\sum_{M_l \in L_{T_i}} m(M_l, T_j) + \sum_{M_k \in L_{T_j}} m(M_k, T_i)}{2}$$

avec :

$$m(M_l, T_j) = \begin{cases} 0 & \text{si } M_l \in T_j \\ \frac{\min_{M_k \in C} DCV(M_l, M_k)}{2} & \text{si } T_j \text{ contient un synonyme de } M_l \quad valeur\_synonyme \\ \min_{M_k \in L_{T_j}} DCV(M_l, M_k) & \text{sinon} \quad valeur\_cooccurrence\_min \end{cases}$$

Notre mesure est une mesure de proximité : les textes sont d'autant plus proches que la mesure est petite.