Etude des résumés en français des rapports de recherche d'un institut d'informatique publiés de 1989 à 1998

A Morin, M. Kerbaol, J.Y. Bansard

Université de Rennes 1 – Campus de Beaulieu – F-35042 Rennes Cedex
e-mail : amorin@irisa.fr, michel.kerbaol@univ-rennes1.fr

Abstract

We focus here on the summaries in french of the internal reports published by INRIA from 1989 to 1998. Our goal is to study the scientific topics of the reports, to bring on the fore a typology of the themes. For that, we use factorial correspondence analysis and the software QNOMIS II which lets us to visualize the results in a friendly way. We can say that the first analyses allow us first to evaluate the quality of the database, then to give some comments about the topics of the reports related to the research themes of the institute. The study is still in process and should be finished in a few weeks.

Résumé

Nous nous intéressons aux rapports internes publiés par l'INRIA de 1989 à 1998. Notre but est de décrire les thèmes abordés et d'en faire une typologie. Nous utilisons pour cela l'AFC et le logiciel QNOMIS II. L'étude est actuellement en cours mais nous pouvons immédiatement évaluer la qualité de la base des rapports, faire émerger les thèmes de publication et les replacer dans les activités générales de l'INRIA. Les premiers résultats sont à la fois attendus et surprenants et conduisent à des interrogations sur la politique de publications internes dans l'institut.

Mots-clés : analyse factorielle des correspondances, classification, mathématiques appliquées, informatique.
correspondence factorial analysis, clustering, applied mathematics, computer science.

1. Introduction

Nous disposons pour l'étude qui suit de la base de données des rapports internes publiés par les différents centres de l'INRIA, Institut National de la Recherche en Informatique et Automatique entre octobre 1989 et décembre 1998. Pendant cette époque, 1976 rapports ont été publiés. En principe, tout rapport est caractérisé par un numéro, une référence, les auteurs, la date de publication (mois, année), la localisation (Rocquencourt, Sophia Antipolis, Rennes, Lorraine, Rhone Alpes), le nombre de pages, un résumé en français et un résumé en anglais, un titre en français et un titre en anglais et enfin une liste de mots-clés dans les deux langues anglais et français, soit 12 paramètres. Cependant ces caractéristiques des rapports n'ont pas toutes été introduites dès le départ et en octobre 1989, seules 8 d'entre elles sont renseignées. En particulier nous ne disposons pas pour les premières années, 1989 et suivantes, des titres des rapports.

Notre objectif est d'étudier d'abord les résumés des rapports en français. Ces rapports sont identifiés par leurs numéro, référence et titre. Nous souhaitons dégager une typologie des thèmes de publication à l'INRIA, d'abord pour l'ensemble de l'institut puis centre par centre. Enfin, nous souhaitons étudier l'évolution des thèmes entre 1989 et 1998. L'étude présentée ici concerne uniquement l'ensemble de l'institut.

Nous utilisons l'AFC pour notre étude. La première étape consiste à constituer le tableau de données avec en ligne le document identifié par son numéro, référence et titre et en colonne les 1000 mots les plus fréquents. Il n'y a aucun traitement linguistique préalable. Cette première étude nous permet d'abord d'évaluer la qualité de la base et d'avoir un premier aperçu des thèmes qui ressortent sur les premiers axes factoriels. Dans l'étude suivante, nous ne conservons que les mots parmi les 1000 dont la contribution à l'inertie est supérieure à la moyenne. Le logiciel utilisé pour notre étude est QNOMIS II, logiciel de cartographie factorielle qui permet de visualiser de façon conviviale les graphiques en utilisant Netscape et qui permet en cliquant sur un point-document de faire le lien avec le résumé original qu'on peut consulter en ligne.

Une première analyse nous a permis de constater que la base des résumés des rapports n'était pas de très bonne qualité. Une seconde analyse après nettoyage et vérification de la base nous permet de retrouver les thèmes développés à l'INRIA tous centres confondus.

2. Première étude

Nous disposons donc de la base des résumés en français et des résumés en anglais des rapports internes de l'INRIA. Entre octobre 1989 et janvier 1999, 1976 rapports ont été publiés dont 1900 rapports de recherche et 76 rapports techniques. Cette base de données est essentiellement utilisée par les documentalistes pour rechercher de l'information. Les consignes de rédaction sont claires, pas de références dans les résumés, pas de formules et pas d'expression en LaTeX. Avant impression et publication, le chef de projet donne uen autorisation signée. Il est clair que ce dernier se soucie surtout du contenu et pas de la forme.

En analyse de données textuelles, la mise en forme et le nettoyage du fichier est une étape indispensable mais très ennuyeuse. Lors de cette première étude, nous nous sommes aperçus que les consignes de rédaction des résumés n'étaient pas suivies et que dans la base il y avait des mots-clés du langage LaTeX, c'est-à-dire avec des caractères inattendus en français ou en anglais. Après nettoyage, nous avons utilisé le logiciel QNOMIS II pour traiter les données.

2.1. QNOMIS II

Le logiciel QNOMIS II développé par M. Kerbaol et J.Y Bansard (Kerbaol et al., 1997) utilise un noyau constitué des programmes d'analyse de données de l'ADDAD. Ce logiciel de cartographie factorielle permet de visualiser de manière interactive les graphiques d'analyse factorielle. Les points du graphique représentant les documents sont liés par des liens hypertexte au contenu du document lui-même ce qui permet d'étudier les raisons des proximités entre documents.

2.2. Résultats

La première analyse est effectuée sur les 997 mots les plus fréquents. Il n'y a eu aucun prétraitement, ni suppression des mots outils, ni lemmatisation. Les mots au singulier et au pluriel sont deux mots différents. A notre grande surprise et malgré le soin pris au nettoyage du fichier, le premier axe factoriel obtenu dans l'analyse des correspondances (Benzecri, 1980; Lebart et al., 1998) des résumés en français des rapports internes de l'INRIA était un axe définis par les mots et les documents en anglais. Les 6 mots dont la contribution à l'inertie était supérieure à deux fois la moyenne étaient *and, in, of, to, is, the,* tandis que 12 documents contribuaient eux aussi à l'inertie de façon significative (plus de deux fois la moyenne). Pour certains de ces 12 documents, il s'agissait d'une inversion des résumés français et anglais ; pour deux autres, il n'y avait

aucun résumé en français. Enfin le résumé français des autres contenaient du yexte en français suivi du résumé anglais sans distinction par un champ spécifique.

Le premier axe n'étant pas très utile pour notre analyse, nous avons examiné avec une certaine attention les axes suivants en sélectionnant les mots dont la contribution était toujours supérieure à deux fois la contribution moyenne. Il nous reste 364 mots qui vérifient cette condition. Rappelons que les projets de recherche de l'INRIA sont organisés par thèmes. Il y en a 4 :

- réseaux et systèmes (1) eux-mêmes répartis en trois sous-programmes :
    1. parallélisme et architecture, (1A)
    2. réseaux, systèmes, évolution ds performances, (1B)
    3. programmation distribuée en temps-réel, (1C)
- génie logiciel et calcul symbolique (2)
    1. sémantique et progammation, (2A)
    2. algorithmique et calcul formel, (2B)
- interaction homme-machine, images, données, connaissances, (3)
    1. bases de données, bases de connaissances, systèmes cognitifs, (3A)
    2. vision, analyse et synthèse d'images, (3B)
- simulation et optimisation des sytèmes complexes (4)
    1. automatique, robotique, signal (4A)
    2. modélisation et calcul scientifique (4B)

Sur le deuxième axe factoriel, il apparaît une opposition du thème 4, simulation et optmisation des systèmes complexes avec une prépondérance de la modélisation et du calcul scientifique sutout de l'analyse numérique au sous-thème du thème 1 parallélisme et architecture tandis que sur le troisième, nous avons une opposition entre le sous-thème du thème 1 réseaux, systèmes, évaluation des performances notamment les files d'attente et processus markoviens et la partie sémantique et programmation du thème 2 notamment spécification et preuves de programmes. Il faut attendre l'axe 4 pour voir apparaitre le thème 3B images qui apparaît assez fréquemment par la suite. Par contre, le thème 3A apparaît seulement sur l'axe 6 par le biais des outils pour documents électroniques.

La distribution des inerties totale des mots est dissymétrique vers la droite. On constate sur de nombreux axes les proximités des mots au singulier et au pluriel par exemple protocole/protocoles, surface/surfaces, etc... qui sont voisins ou confondus.

3. Deuxième étude

Dans cette étude, nous nettoyons le fichier en supprimant tous les codes LaTeX, les mots de longueur inférieure à 3 et en corrigeant les inversions des résumés français et anglais. Il nous reste 1940 documents. Nous sélectionnons pour la première analyse 967 mots. Dans ce cas, il n'y a pas de suppressions des mots outils, ni de lemmatisation. A l'issue de cette première analyse, nous sélectionnons les mots dont la contribution à l'inertie des axes est supérieure à deux fois la moyenne. Nous demandons 60 axes. Il nous reste 420 mots. Nous sommes surpris par le nombre de mots sélectionnés.

La distribution des inerties totales des 967 mots est plutôt symétrique de moyenne 0.98 avec un écart-type de 0.49 : en fait 43% des mots ont une contribution supérieure à la contribution moyenne ce qui laisserait supposer des documents assez homogènes au niveau du vocabulaire.

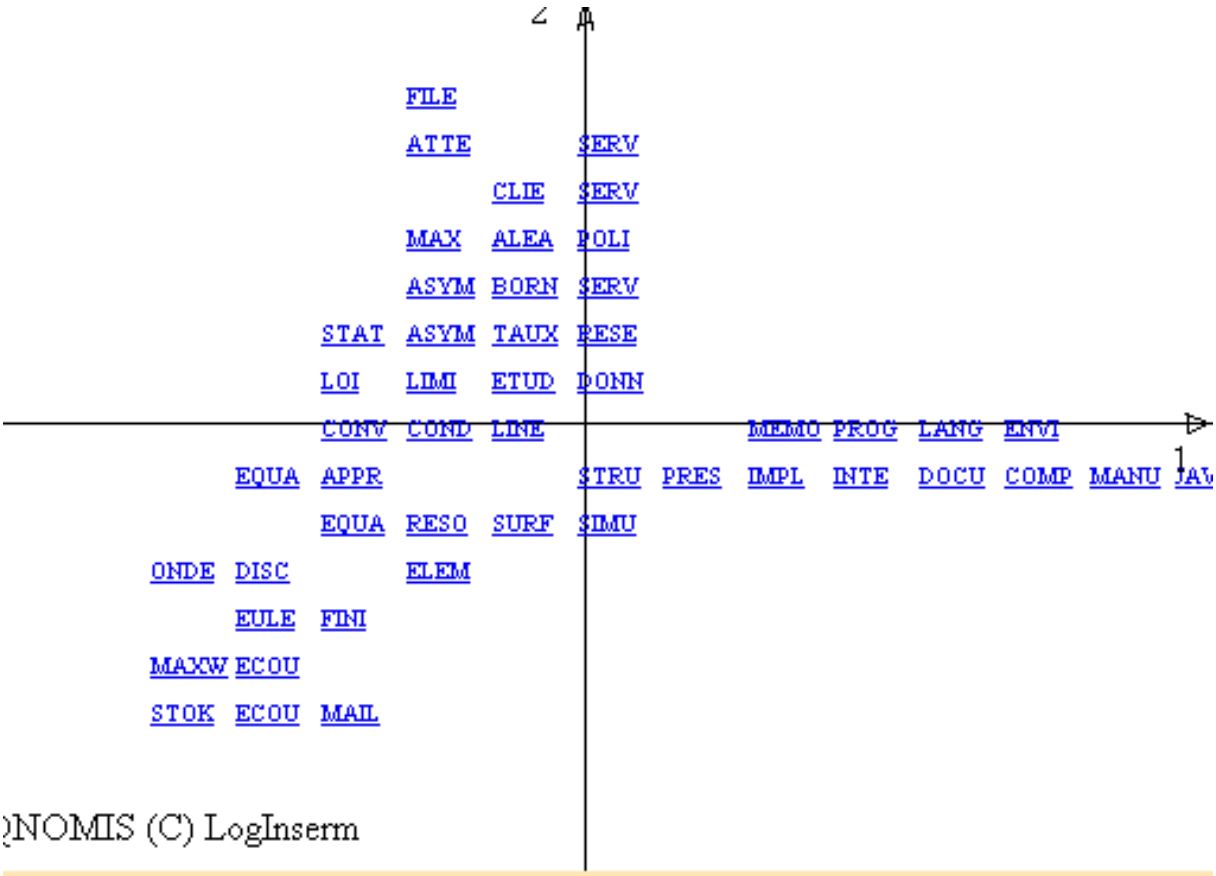
L'élimination ou la correction des résumés en anglais nous permet de "zoomer" sur les plans

factoriels précédents numérotés 2-3 ou plus . On retrouve sur le premier plan (cf. graphique) les thèmes 1, 2 et 4 tout comme dans l'analyse précédente. Le thème 3 qu'il s'agisse d'images ou de documents apparait sur les plans suivants.

Une des difficultés de l'interprétation, si on la veut plus fine ets l'absence de réference à un projet particulier. Compte tenu des mots et du document, il faut replacer le document dans son thème. Par ailleurs, Qnomis II permet pour un mot donné d'afficher les axes auxquels ce mot contribue dans l'ordre des contributions décroissantes. Cela permet notamment de rechercher à partir de mots spécifiques le thème 3A qu'on estime peu ou mal représenté. Il y a des mots comme manuel en général proche de documents qui correspond souvent à des rapports qui sont des manuels d'utilsation de tels ou tels logiciels. Ce ne sont pas de srapports de recherche mais plutôt des rapports tecniques. malheureusement lorsqu'on examine les rapports tecniques, il est difficile de leur trouver une caractériqtique commune. certains sont des manuels d'utilisation , d'autres des documents de cours. Quelques sont très techniques, description de systèmes ett les autres ressemblent à des rapports de recherche. Il est extrêment difficile de retrouver les règles de publication à partir de ces documents.

4. Conclusion

Cette étude qui n'est pas terminée a apporté plus d'interrogations qu'elle n'en a résolu. Bien que les rapports internes de l'INRIA doivent refléter l'activité de l'institut, on a l'impression que le soin apporté à la classification du document,(rapport de recherche, rapport technique) , au résumé ou au choix des mots-clés n'est pas tout à fait à la hauteur de l'enjeu. Nous avons eu aussi l'occasion de constater chez les documentalistes une sorte de résignation à propos de nos remarques concernant la mauvaise qualité des résumés (pas scientifique mais rédactionnelle) : on ne peut pas et on ne réussira pas à imposer au chercher les règles concernant la publication. Pourtant, ce sera le prix à payer pour avoir une base de données performante et exploitable par les documentalistes.



Références

Benzecri J.-P. (1980). *Pratique de l'analyse des données*, volume Tome 1 : analyse des correspondances. Dunod.

Kerbaol M., Bansard J., and Favier L. (1997). Sélection de la bibliographie de "maladies rares" : une approche expérimentale. In *4ième séminaire d'Obernai DIDDOC INSERM*.

Lebart L., Salem A., and Berry L. (1998). *Exploring textual data*. Kluwer Academic.