

Expériences de classification d'une collection de documents XML de structure homogène

- Thierry Despeyroux
- Yves Lechevallier
- Brigitte Trousse
- Anne-Marie Vercoustre

Projet Axis

Inria, Rocquencourt et Sophia Antipolis

E_mail: prénom.nom@inria.fr

Le Rapport d'Activité Scientifique de l'Inria

Project-Team : axis - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.inria.fr/rapportsactivite/RA2003/axis2003/axis_tf.html Search

Home Bookmarks

AXIS

Team

Overall Objectives

- Objectives

Scientific Foundations

- Semantics and Design of Hypertext Information Systems
- Usage Mining: Applying KDD to Usage Data
- Adaptive Recommender Systems (Mihai)
- Case-Based

User-Centered Design, Improvement and Analysis of Information Systems

AXIS

2003 research project activity reports

Sophia Antipolis

Theme: 3a

Presentation of the project - Activity report in PostScript or PDF format

EGC 2005

Une présentation homogène

Project-Team : axis - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.inria.fr/rapportsactivite/RA2003/axis2003/axis_tf.html Search

Home Bookmarks

AXIS

Team

Overall Objectives

- Objectives

Scientific Foundations

- Semantics and Design of Hypertext Information Systems
- Usage Mining: Applying KDD to Usage Data
- Adaptive Recommender Systems (Mihai)
- Case-Based

- Team
- Overall Objectives
 - Objectives
- Scientific Foundations
 - Semantics and Design of Hypertext Information Systems
 - Usage Mining: Applying KDD to Usage Data
 - Adaptive Recommender Systems (Mihai)
 - Case-Based Reasoning
- Application Domains
 - Panorama overview
- Software
 - Introduction
 - CLF - ``Computer Language Factory''
 - Clustering ToolBox
 - CBR*Tools - Object-oriented Framework for Case-Based Reasoning
 - Broadway*Tools - Generator of Adaptative Recommender Systems
 - Broadway-Web - Personalized Supporting Web Browsing
- New Results
 - Data Transformation and Knowledge Representation
 - Data Mining Methods
 - Viewpoint Management in KDD

Le RA en chiffres

- 146 fichiers
- 229 000 lignes
- 14,8 M octets de données
- Une DTD unique
- Des sections optionnelles
- Le style et le contenu sont libres

Classement en Thèmes (2003)



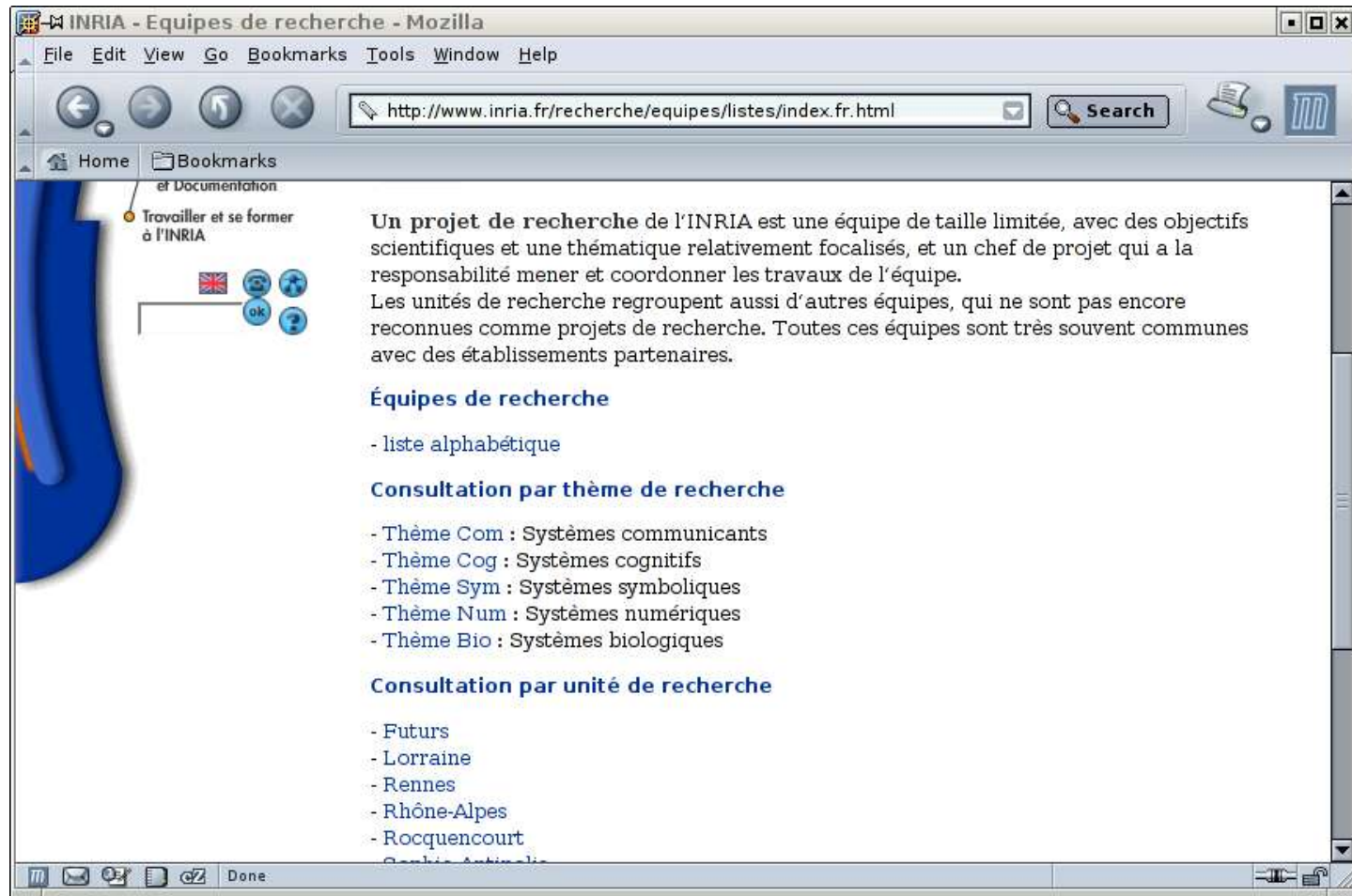
The screenshot shows a Mozilla browser window with the title "INRIA - 2003 Scientific Activity Reports - Mozilla". The address bar contains the URL "http://www.inria.fr/rapportsactivite/RA2003/index.html". The page content includes the INRIA logo, a "Back to homepage" link, and an "INDEX" link. The main heading is "INRIA - Scientific Activity Report 2003". Below this, there is a list of four themes:

- theme 1 : Networks and Systems
- theme 2 : Software Engineering and Symbolic Computing
- theme 3 : Human-Computer Interaction, Images Processing, Data Management, Knowledge Systems
- theme 4 : Simulation and Optimization of Complex Systems

Below the list, the text "In alphabetic order" is displayed. At the bottom of the page, three items are listed:

- A3 - Advanced Analysis to Code Optimization
- ACACIA - Knowledge Acquisition for Aided Design Through Agent Interaction
- ACES - Ambient Computing and Embedded Systems

Classement en Thèmes (2004)



Le Problème

- Une présentation sous forme de thèmes de recherche
- Qui peut varier dans le temps
- Qui n'est pas neutre politiquement

- Existe-t-il des regroupements naturels ?
- Quel est le rôle des différentes parties ?

Méthodologie

1. Sélectionner des parties pertinentes en utilisant la **structure XML**
2. Sélectionner des mots en utilisant un outils de **typage syntaxique** (TreeTagger) et en les lemmatisant
3. Regrouper les documents en un ensemble de classes disjointes
4. Evaluer ces regroupements

Les diverses expériences

- K-F: mots-clefs de la section *fondations*
- K-all: tous les mots-clefs
- T-P: texte de la section *présentation*
- T-PF: texte des sections *présentation* et *fondations*
- T-C: noms de conférences, workshops, congrès etc. dans la *bibliographie*

TreeTagger

XML

Tree Tagger

A3	presentation	a3	JJ	<unknown>
A3	presentation	designs	NNS	design
A3	presentation	methods	NNS	method
A3	presentation	and	CC	and
A3	presentation	tools	NNS	tool
A3	presentation	used	VVN	use
A3	presentation	by	IN	by
A3	presentation	compilers	NNS	compiler
A3	presentation	or	CC	or
A3	presentation	users	NNS	user
A3	presentation	for	IN	for
A3	presentation	code	NN	code
A3	presentation	analysis	NN	analysis

EGC 2005

Méthode de Classification

L'objectif de cette troisième étape est de regrouper les documents en un ensemble de classes disjointes à partir des **vocabulaires** issus des cinq expériences.

Pour réaliser ce classement nous utilisons une **méthode de partitionnement de type Nuées Dynamiques** où la distance est basée sur les fréquences des mots du vocabulaire choisi.

Le principe de l'algorithme est proche de l'algorithme des k-means.

Experience K-F-a: liste des mots-clefs représentatifs des classes

Classe 1: 3d approximation, computer, differential, environment , modeling, processing , programming , vision

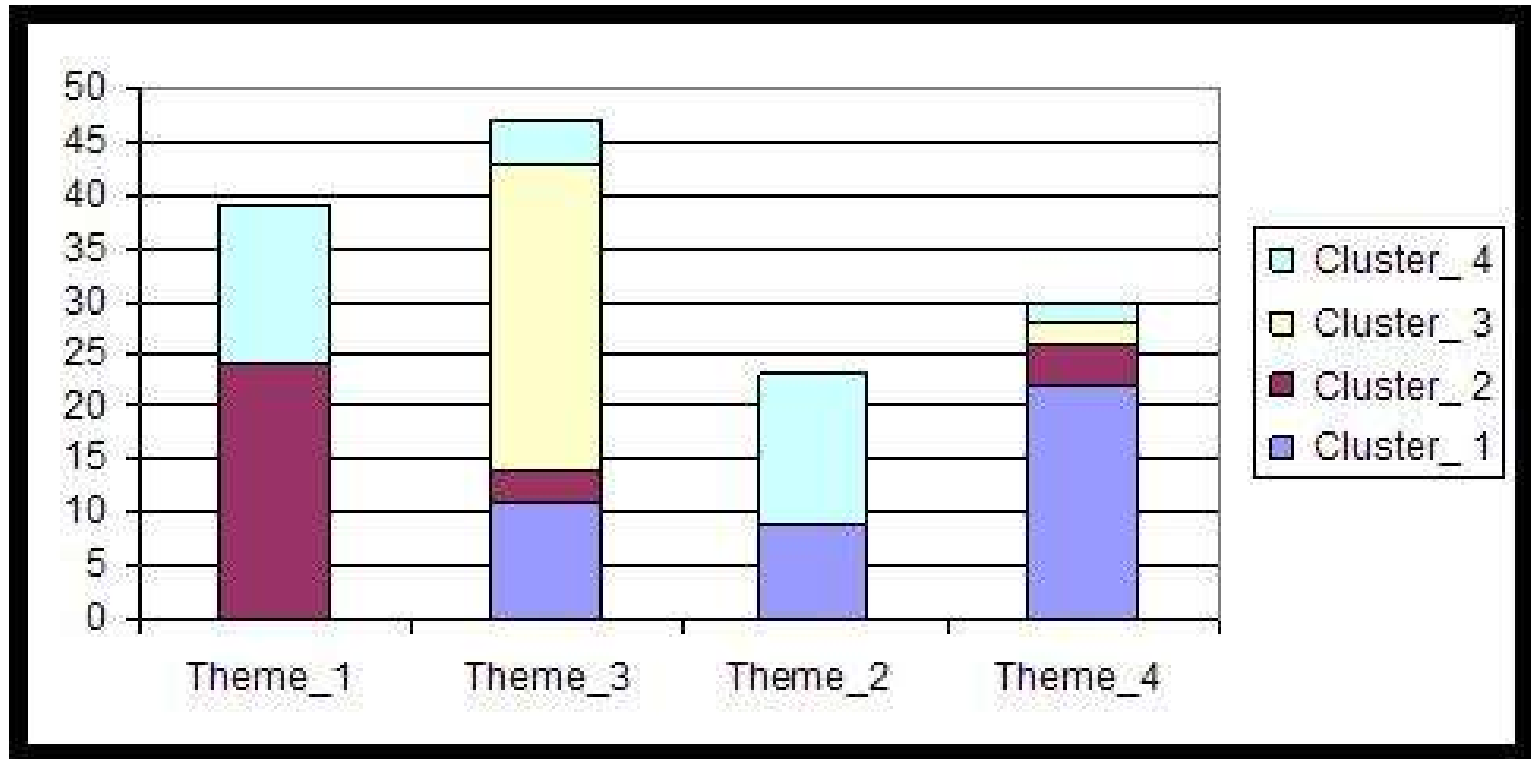
Classe 2 : computing, equation, grid, problem, transformation

Classe 3 : code, design, event, network, processor, time, traffic

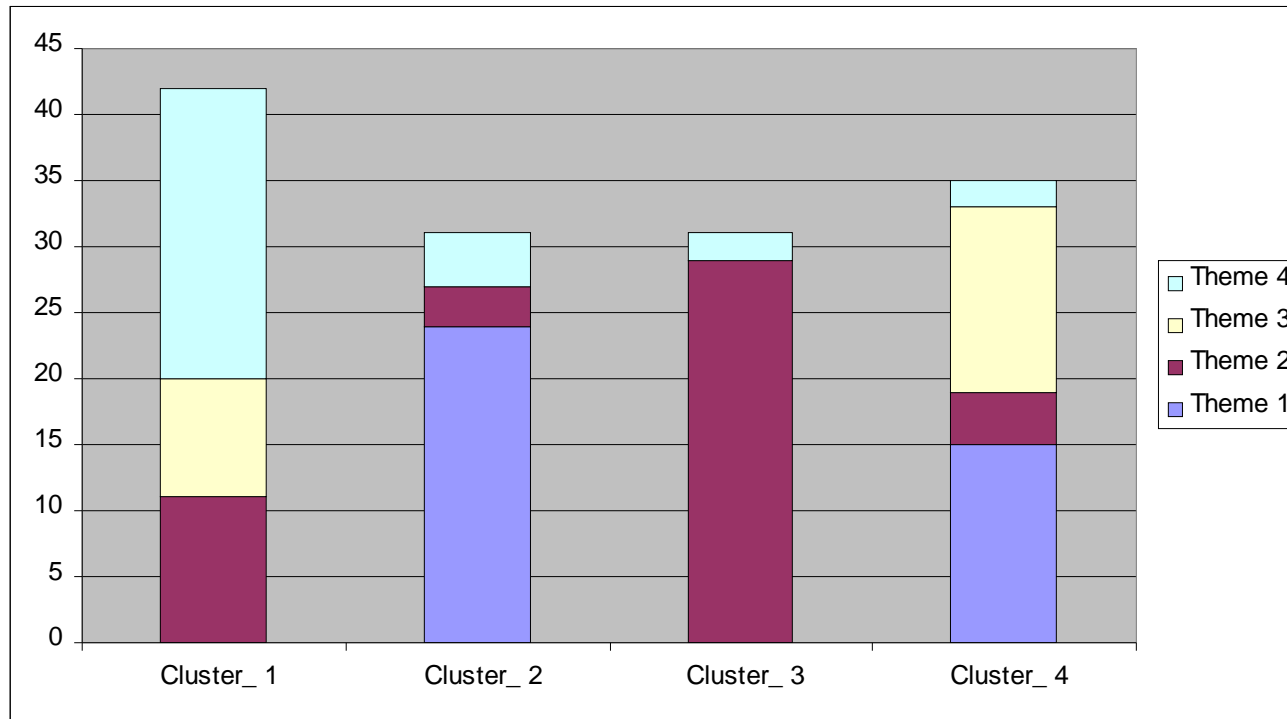
Classe 4 : calculus, database, datum, image, indexing, information, integration, knowledge, logic, mining, pattern, recognition, user, web

A chaque classe est associé les mots les plus représentatifs de cette classe. Ces mots peuvent être interprétés comme des **résumés** de ces classes.

Répartition des classes en fonction des thèmes de 2003



Répartition des thèmes de 2003 en fonction des classes



Partition des projets

Thèmes	Cluster_1	Cluster_2	Cluster_3	Cluster_4
1a	Apache Runtime	Tropics Regal	R2D2 Arles Caps POPS	
1b		AlGorille Gyroweb	Reso armor Mascotte aces	
1c		ADEPT DaRT	Ostre Pop Art Triskell VerTeCs	
2a	Lande PROTHEO SECSI calligramme	COMPOSE		LogiCal
2b	Cafe LEMME	Adage Algo		

Partition des projets

Thèmes	Cluster_1	Cluster_2	Cluster_3	Cluster_4
3a	i3D DREAM METISS Parole PRIMA SIGNES Atoll	MOSTRARE WAM	cordial	TEXMEX ACACIA Cortex Symbiose AXIS Orpailleur Gemo Orion ATLAS Smis
3b	Mirages ALCOVE REVES siames	VISTA artis	TEMICS macsi	Air2 Imedia
4a	icare BIPOP Miaou Imara COMORE	Fractales NUMOPT IS2 corida	sigma2	
4b		Caiman Calvi Smash Opale ALADIN	sagep	sagep

Méthode d'évaluation externe

L'évaluation de la qualité des classes générées par la méthode de classification est basée sur sa comparaison des classes ainsi obtenues avec les deux listes de thèmes utilisés par l'Inria.

n_{ij} est le nombre de projets de recherche INRIA ayant leurs rapports de recherche classés dans la classe U_i et ayant été affectés au groupe C_k du thème k .

n_i est le nombre de rapports de recherche mis dans la classe U_i ,

$n_{.k}$ est le nombre de projets affectés au groupe C_k du thème k ,

n est le nombre de projets analysés.

Deux mesures

La **F-measure** proposée par (Larsen and Aone, 1999) combine les mesures de **précision** et de **rappel** entre U_i et C_k .

La mesure de rappel est définie par $R(i,k)=n_{ik}/n_i$.

La mesure de précision est définie par $P(i,k)=n_{ik}/n_{.k}$.

La F-measure entre la partition a priori U en K groupes et la partition P des projets INRIA obtenue par la méthode de classification est :

$$F = \sum_{k=1}^K (n_{.k}/n) \max_j (2 \cdot R(i, j) \cdot P(i, j) / (R(i, j) + P(i, j)))$$

L'index **corrected Rand** (CR) proposé par (Hubert and Arabie (1985)) pour comparer deux partitions.

Mesure de validité externe résultats

		Thèmes2003		Sous themes 2003		Thèmes2004	
Exp.	K	F	Rand	F	Rand	F	Rand
K-F-a	4	0.53	0.14	0.38	0.09	0.46	0.11
K-F-b	5	0.44	0.05	0.35	0.06	0.37	0.03
K-F-c	9	0.42	0.10	0.37	0.08	0.43	0.12
K-all-a	4	0.52	0.17	0.36	0.09	0.47	0.15
K-all-b	5	0.53	0.17	0.37	0.10	0.54	0.22
K-all-c	9	0.46	0.13	0.40	0.12	0.38	0.10
T-P-a	4	0.55	0.19	0.40	0.14	0.50	0.19
T-P-b	5	0.45	0.11	0.42	0.12	0.47	0.15
T-P-c	9	0.44	0.11	0.45	0.16	0.44	0.14
T-PF-a	4	0.66	0.32	0.49	0.27	0.50	0.21
T-PF-b	5	0.56	0.22	0.43	0.18	0.51	0.20
T-PF-c	9	0.48	0.22	0.55	0.29	0.46	0.19
T-C-a	4	0.51	0.15	0.39	0.15	0.50	0.21
T-C-b	5	0.44	0.18	0.45	0.24	0.47	0.17
T-C-c	9	0.45	0.13	0.47	0.21	0.45	0.15

Conclusion

- Combinaison d'une sélection par la structure et par le contenu linguistique
- Évaluation de la classification par rapport à une typologie existante
- La qualité de la classification dépend fortement du choix des parties des rapports d'activité prises en compte

Futur :

- Évolution de ces classes dans le temps
- Expérimentation sur d'autres collections