INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team KIWI

Knowledge, Information and Web Intelligence

Nancy - Grand Est

THEME COG

Table of contents

KIWI is a project of University Nancy 2 through LORIA laboratory (UMR 7503). The team has been created on January the 1st, 2008. For more details, we invite the reader to consult the team web site at http://kiwi.loria.fr/.

1. Team

Faculty Member

Anne Boyer [Professor, Nancy2 University, HdR]
Armelle Brun [Assistant professor, Nancy2 University]
Azim Roussanaly [Assistant professor, Nancy2 University]
Alain Lelu [Professor, Université Franche Comté, HdR]

PhD Student

Cedric Bernier [scholarship CIFRE from february, Nancy2 University]
Geoffray Bonnin [scholarship Region Lorraine, Nancy2 University]
Sylvain Castagnos [ATER, until november, Nancy2 University]
Ilham Esslimani [scholarship Region Lorraine, Nancy2 University]
Le Hong Phuong [scholarship Region Lorraine, Nancy2 University]

Administrative Assistant

Céline Simon [TR, INRIA]

2. Overall Objectives

2.1. Research Objectives

The objective of the researches conducted in the KIWI team is the increase of the quality of online services (intranets, numerical databases, information portals, etc.) brought to an identified or non-identified active user.

More specifically, the goal is to improve the quality of the interactions between the general audience and systems of research and access to information. There are several possible approaches to assist the active user: adaptive interfaces to facilitate the exploration and the searches on the Web, systems relying on social navigation, sites providing personalized content, statistical tools suggesting keywords for improving searches, etc. Our approach consists in providing each user with items likely to interest him/her. Contrary to the personalized content, this solution does not require to adapt resources to the potential readers. Each item has to be proposed to concerned persons by using push-and-pull techniques.

2.2. Highlights

The KIWI team has been created on the 1st, Januray, 2008.

3. Scientific Foundations

3.1. Introduction

While navigating on the Internet, the user expresses more or less complex requests by using keywords or logical connectors, which is a difficulty for the large audience.

Requests are usually processed independently of the user and the context. The criterions taken into account are the resources and their content, their popularity and availability. The only way a user can have his interests or preferences taken into account is to explicitely express them in the requests, which is irrealistic.

Moreover, this approach requires that the resources are tagged or labelled to establish a link between the requests and the resources. However, automatically labelling resources with pertinent and contextualized keywords is a complex problem that does not have any generic solution at the moment. Furthermore, the labelling can evolve along time and depends on both the "point of view" of the users that label resources, and the context.

Moreover, researches based on keywords and logical connectors do not guarantee anymore the identification of the most pertinent resources at a fixed time, as the number of results of the requests is huge. For example, at the beginning of 2007, 5 billions results were obtained when requesting "news" on Google and about 400 millions results were obtained when requesting "news New-York". Furthermore, one keyword may reflect various meanings according various users; at the opposite some different requests may correspond to the same search. Finally, keywords-based research is not efficient: the well-known recall and precision criteria are below 1% and are thus unsatisfactory.

Another approach consists in constructing a training relationship, *i.e*. a privileged relationship between a service and a user. This relationship permanently integrates new information in order to increase the satisfaction of the user. The approach we adopt is mainly this one and can thus be viewed as following:

- Training of the models of users behavior, based on the observation of the interactions between the users and the system
- Detection of the behavior of the active user based on his/her interactions with the system
- Planning of the actions the system will do, based on the observed behavior of the active user, in order to provide a service dedicated and adapted to the active user expectations

This approach is thus a usage-based approach, it requires the availability of a large usage corpus (or a trace corpus) to construct accurate models. Obviously, the system cannot wait for collecting a sufficiently large set of observations about a user to construct a model and then adapting the system to him, along the model. We propose here to take advantage of the information collected about other users. The training will thus be **collective: the knowledge about some users is exploited for the others**.

Reinforcement is a relevant training technique in this context. It enables the system to incrementally learn about the user, based on the principle of "punishment/reward". At the opposite of classical approaches, the reinforcement signal also comes from other users.

The usage-based approach can be viewed as complementary to the content-based approach (using tags and keywords); we are thus also interested in the content based approach and hybrid approaches, that mix usage-based and content-based approaches.

The quantity of information to consider quickly becomes too huge (in terms of space complexity and time processing), thus a decentralized approach enables to lower the complexity of processing.

The KIWI project is thus related to the user behavior modelisation, based on traces of usage, on content and on collective reinforcement learning, and decentralized training.

3.2. Usage Based Learning

The usage-based approach does not take into account the content of a resource, it focuses only on the interest of the user for this resource. The information available about a resource is only its *id*: no information about its content, date of creation, author, tags, etc. is known. In the same way, no information *a priori* about the user is available. The measure of interest is more subjective than any measure depending on the content of resources, it is directly linked to satisfaction.

The usage-based approach consists in exploiting the large number of traces the users leave deliberately or not, during their interactions with a service. These traces can be explicit (ratings, tags, annotations) or implicit (clickstream, logs, consultation histories). These traces are then used to infer information about the interests of users.

3.2.1. Objectives

No model *a priori* exists, and no model can exist, about users, about resources, and about the interaction between users and resources. This lack is mainly due to the subjective dimension of the problem we face: the interest of a user for a document, the satisfaction of the user, the polysemy of actions, etc. Moreover, the set of users and the set of resources are in essence dynamic.

The objective is thus to develop a model of interest of the user for a resource, based on the traces (explicit or implicit) of interactions of the user with the service (and the resources).

3.2.2. Approach

The quantity of available data leads us to exploit them by using statistical methods, in order to build, by using automatic training, the pertinence function of a resource for a user, according to his context.

To overcome the lack of information available about a user at a given moment, we choose to have a **collaborative approach** that consists in exploiting the data available about other users. These other users are the ones that have a behavior similar to the user we aim at modeling. Thus, our work is in the frame of the WEB2 (social) and the WEB3 (intelligent).

To develop the model of interaction between the user and the system, we are interested in the two standard points of view: user-based approach and item-based approach.

3.3. Hybrid Learning

The principle of hybrid recommendation systems is based on the joint using of collaborative filtering and content analysis of resources. The content related knowledges can be explicit (e.g. for movies application, they can be the style, the cast, the director...) or implicit (e.g. for web pages application: the links, the size...)

In the perspective of increasing efficiency by reducing data space, following the example of the model oriented approach for the collaborative filtering, clustering methods based on the content features may be adapted.

One way of investigation of our team is the adaptation of formal clustering methods with the aim of taking into account the content of resources.

Another way of investigation is the using of natural langage processing methods and document indexing fields tools and algorithms in order to identify the most significant features and to extract them from the text part of resources content.

4. Application Domains

4.1. Application Domains

The new communication technologies are predominant in our professional and personal environment. They enable a generalized diffusion, instantaneous and at every place of the information. The development of the Internet is such that every day there are more and more users, more and more online services and more and more content.

Internet is now accessed by users less and less aware of technology and more and more demanding, which is a problem we have to face. Moreover, the satisfaction of a user depends on his expectations, his knowledge, abilities or interests and also on his context. Thus, a service cannot anymore be developed according to predefined scenari or *a priori* models of the world.

The researches conducted in the KIWI team are thus applied to the domain of personalisation of services, in the frame of

- Intranets
- Web navigation
- E-commerce
- etc.

5. Software

5.1. FRAC+

5.1.1. FRAC+

is a software developed in JAVA which fits with client/server architectures. Its goal is to provide a personalizing service within enterprise portals and e-commerce applications. The client part includes a module for user modeling and a recommender engine. The server part contains a clustering collaborative filtering algorithm allowing the system to build virtual communities of interests. This enables to acquire experience from similar population and to bypass the problems of missing data in individual preference models, thus improving the recommendation computations on the client side. This software has been successfully integrated in real industrial contexts: the `Casablanca` website broadcasting service of the company SES ASTRA (within the framework of the ESA Sat'n'Surf project), the intranet portal of the technological foresight department of Crédit Agricole S.A. (Paris, France). It has also been used through a collaboration with the Artificial Intelligence Laboratory at EPFL (Lausanne, Switzerland). At last, this software has been the subject of a technological transfer with the company Sailendra S.A.S. (Nancy, France).

- *Availability*: distributed by Sailendra S.A.S.
- *Contributors*: Sylvain Castagnos, Anne Boyer
- *Contact*: sylvain.castagnos@epfl.ch, contact-sailendra@loria.fr

5.2. SOFOS

5.2.1. SofoS

is an experimental peer-to-peer documentary platform allowing people to share items in various formats (textual, video, audio, websites, etc.) and to do some researches in order to find new contents available on the platform. This software has been developed in JAVA and relies on the `JXTA`[1] open-source toolkit. `SofoS` includes a powerful recommender system and enables the KIWI team to test and validate new filtering algorithms. This software already includes the highly-scalable AURA collaborative filtering algorithm which provides very accurate recommendations based on implicit and explicit criteria in real time. This platform is currently being industrialized within the framework of the `CIA` Project (Cérès) and is intended for detecting unusual behaviors and malicious users whose goal is to influence other users and to provide false recommendations.

- *Availability*: Waly secured server (LORIA)
- *Contributors*: Sylvain Castagnos, Anne Boyer
- *Contact*: anne.boyer@loria.fr

5.3. Casaweb

CasaWeb is a Collaborative Filtering software developped in Java. The objective of CasaWeb is to construct users profiles and generate predictions about preferences of users by using exclusively usage traces and exploiting navigational patterns. These traces are related to navigational activities of users collected from web server log files. CasaWeb does not need any explicit ratings as input, it can estimate users ratings from the usage traces by considering implicit criteria about each consulted resource.

- *Availability*: Not distributed
- *Contributors*: Ilham Esslimani, Armelle Brun, Anne Boyer
- *Contact*: anne.boyer@loria.fr

[1] https://jxta.dev.java.net/

5.4. vnToolkit

vnToolkit is an Eclipse Rich Client Application which integrates some tools for automatic processing of Vietnamese texts. It provides a general framework and GUI for hosting several tools for natural language processing of Vietnamese. At the moment of this writing, vnToolkit compises of the following tools:

1) vnTokenizer: an automatic tokenizer of Vietnamese texts which deploys automata techniques and statistical approach. 2) vnSentSegmenter: an automatic sentence boundary detector of Vietnamese texts which uses a maximum entropy approach. 3) vnTagger: an automatic part-of-speech tagger of Vietnamese texts which uses a maximum entropy approach.

- *Availability*: http://www.loria.fr/~lehong/projects.php, GNU General Public License.
- *Contributor*: Le Hong Phuong
- *Contact*: phuonglh@gmail.com, lehong@loria.fr

5.5. SynAF

SynAF annotator is a graphical syntactic annotation framework for building treebanks. The tool is developed in the Java programming language. It is bundled as an Eclipse Rich Client Application and distributed under the GNU General Public License.

- *Availability*: http://www.loria.fr/~lehong/projects.php, GNU General Public License.
- *Contributor*: Le Hong Phuong
- *Contact*: phuonglh@gmail.com, lehong@loria.fr

6. New Results

6.1. Introduction

A. Roussanaly and P. Le Hong are formerly members of the TALARIS team. They continued with their ongoing project in the field of natural langage processing, especially on syntax and parsing (see Bibliography section).

6.2. Usage Based Recommender Systems

The results reached this year in our new team are mainly in the frame of usage-based recommender systems.

6.2.1. Generic preference modeling function dealing with Privacy

Participants: Sylvain Castagnos, Anne Boyer.

To supply the active user with his/her concerns, we first have to build his/her model of preferences by collecting data about his/her activities. This approach is based on an analysis of usage. The data collection method can be either explicit, either implicit. In the first case, personalization relies on information explicitly provided by users, such as ratings or demographics. For example, users may rate a sample of items in order to receive suggestion of new items that may interest them. On the opposite, personalization based on implicit data collection infers unknown preferences of a user from his/her browsing history, purchase history, search queries, etc.

Our contribution [5] consists in proposing a generic function for modeling preferences of users. This function can take into account a large number of criteria such as the explicit ratings, the frequency and duration of consultations, the number of keywords provided by users, the actions of printing or saving items, etc. These criteria can be chosen according to the context of use of the system (available pieces of data). The modeling function is extremely flexible by being able to integrate as many new criteria as we want. This function then returns numerical user profiles under the form of estimated ratings. This measures the potential interest of each user for each item.

We particularly paid attention to the privacy concern [21] and propose a good compromise between accuracy of predictions and privacy of users. We define a generic procedure to comply with these goals (preferences collected for a limited duration and exploited anonymously, pieces of data transformed into a less intrusive form than log files, low level of involvement needed from users, transparent process, etc.).

6.2.2. Item-based approach

6.2.2.1. Reinforcement Rules for Item-Based Recommender Systems

Participants: Sylvain Castagnos, Armelle Brun, Anne Boyer.

Collaborative Filtering algorithms provide personalization by exploiting the knowledge of a similar population and predicting future interests of a given user (called "active user") as regards to his/her known preferences. In practical terms, this kind of algorithms is broken down into 3 parts. First, the system needs to collect data about all users as explicit and/or implicit ratings. Second, this data is used to infer predictions, that is to say to estimate the votes that the active user would have assigned on unrated items. Finally, the recommender system suggests to the active user items with the highest estimated values. The aim of this work was to improve the second step for Item-Based systems where classes of similar items are built by computing each pairwise correlations.

We assume that, in some cases, pairwise similarities may be insufficient to explain the interest of a user for an item. Guided by this hypothesis, we propose a new model, called `RIBA` [12], which evaluates similarities of triplets, rather than pairs of items. To illustrate this statement, we can consider three items i_k="Cinderella", i_t="Scary Movie", and i_w="Shrek". A user may have liked i_k which is a fairytale without appreciating i_w. At the same time, a user who enjoys the horror film parody i_t should probably rate lowly i_w. However, a film goer who likes both fairy tales and parodies will take fun when watching Shrek.

Our main contribution consists in automatically detecting non-trivial associations between items in order to make the recommender engine more accurate. These associations take the form of reinforcement rules. These rules are determined by exploiting probabilistic skewnesses in triplets of items, and are then used to refine predictions of the system. Our approach only requires explicit ratings from users and doesn't need any information about items.

Our experiments have shown that some predictions are enough accurate using pairwise similarities and that some other predictions need to be refined with our association rules. Thus, we propose in [11] a prediction confidence metric which estimates the expected accuracy of each prediction based on pairwise similarities, and allows the system to automatically decide if reinforcement rules are adapted to the situation.

Despite a computational process which is still time-consuming, our model shows an improvement from 6 to 8% as regards the HMAE accuracy measure and constitute a first step to build scalable tuple-based recommender systems.

6.2.2.2. Order Dependent Recommendation Models

Participants: Geoffray Bonnin, Armelle Brun, Anne Boyer.

We are interested in the exploitation of statistical natural language processing models to enhance Collaborative Filtering in the frame of Web ressources. We propose to use an enhancement of n-gram models called skipping [26], which allows to take into account discontiguous sequences within a fixed size window. This thus allows a longer history than contiguous n-gram models, while having the same low complexity. Increasing the history size makes the model provide a better accuracy and a better coverage than a contiguous n-gram model, as well as robustness to noise and the ability to take into account multiple parallel navigations. These last features were empirically validated using artificial corpora containing an increasing amount of noise. As there are many variants to apply for skipping, we tested three of them: a low complexity state-of-the art variant, a higher complexity variant that includes all possible combinations within the window, and a third variant we designed especially for Web navigation which provided results similar to the second variant while having the same complexity than the first one. This work has been published in [7].

Further work has been led, in which a new feature has been evaluated: weighting the occurences according to the distances between the elements of the skipped n-grams. Four weighting schemes were evaluated among which a new one was designed. Evaluations were done on an Intranet browsing dataset provided by the Crédit Agricole, a famous French bank. Results, that have been published in [8] confirms previous results on synthetic data, and show our weighting scheme provides the best results.

In a following work [9], we exploited one particular feature of our algorithm to reduce runtime performance. Indeed, the way we combine the probabilies when considering all combinations whithin the fixed size window during recommendations is done according to an anytime scheme. This scheme allows us to stop the computations of recommendations after a certain amount of loops. Evaluations shows that the quality of the recommendations increase in an asymptotic way, allowing a good tradeoff between computation time and precision.

6.2.3. User-based approach

6.2.3.1. Navigation-based Recommendation Models

Participants: Ilham Esslimani, Armelle Brun, Anne Boyer.

We have proposed a recommender system based on a new navigational technique that takes into account common navigational patterns, in order to compute correlations between users and select neighborhoods, without using any rating data. The originality of this recommendation system is the use of navigational patterns, in place of ratings, to compute correlations between users. The performance of this recommender system is tested without and by combining predictions of both navigational based technique and classical collaborative filtering, in terms of accuracy and robustness (by using Movielens Datasets). The experimentation put forward the impact of the navigational based technique on the performance of the recommender system in terms of precision and robustness. The tests show that the more the navigational based technique is involved in the recommendation process, the more high predictions are accurate ([6]) ([14]).

6.2.3.2. Clustering in Navigation-based Recommendation Models

Participants: Ilham Esslimani, Armelle Brun, Anne Boyer.

We also studied the exploitation of navigational patterns (mentionned above) within the frame of a new approach of Collaborative Filtering. This approach uses usage traces to estimate ratings in order to construct a rating similarity matrix of users. By using this matrix, users are clustered based on a partitioning algorithm called PAM (Partitioning Arouns Medoids). Then, navigational correlations between users are computed within these clusters, by taking into account the longest common sub-sequence between pairs of users. The navigational correlations are based on positive sequences that integrate only the preferred resources of users in order to improve the time processing. The performance of the proposed approach is evaluated in terms of accuracy and time processing on a real usage dataset (extracted from Crédit Agricole Banking Group, in particular the usage data relating to the Department of Strategies and Technology Watch. The experimentation shows that this approach highly improves the accuracy of predictions. Moreover, the use of clustering and positive sequences contributes to an important reduction of dimensionality and time processing required for computing the navigational correlations.

6.2.3.3. Recommendations for groups

Participants: Cedric Bernier, Armelle Brun, Anne Boyer.

We started this year studying the recommendations for groups. There are two ways to make a recommendation for groups, by creating a group profile then compute recommendations or by using the user profiles of the users of the group and compute a recommendation, based on all these user profiles. We are investigating both approaches and testing them on applications such as video clips selection in public area or movie selection at home.

6.2.3.4. Statistically valid graphs for usage mining

Participant: Alain Lelu.

Relations between variables in a datatable, or between individuals, are the building blocks of a bunch of further processings, especially for neighborhood-based collaborative filtering: I have focussed on the applications of the statistical validation of such relations, using the Tournebool randomization test [25] for setting up the graph of valid variable relations in a Reuters' test corpus [18] and exploring its properties. An extension to higher-order relations has also been proposed [10]. My "valid graph" approach has been implemented and tested positively in the framework of the GERMEN project (INIST collaboration) towards an incremental representation of a textual database records stream [27].

6.3. Hybrid Recommender Systems

6.3.1. A metrics for directional content-data

Participant: Alain Lelu.

Directional data are collections of large sparse vectors, which association is independant from their norms. These data generally encounter in the context of text databases, gene expression analyses, and collaborative filtering [24]. My first research axis consists in new factorization methods for these directional data, and reconstruction of missing values. It has been established by the cited authors and [28] that the usual kmeans method using euclidean distance performs much worse than spherical kmeans for directional data. My Axial kmeans method is a variant of spherical kmeans using Hellinger metrics [22]. In the sequel of a first large scale application of AKM to text mining on a 170 000 abstracts bibliographic database [29], a collaboration with the Observatoire des Sciences et des Techniques (Paris), I have implemented a multifaceted comparison along two different views on the same big document collection: the first using citation information, the second using lexical elements in the titles and abstracts [19]. An article for the WoS-indexed Journal of Informetrics (L. Egghe ed.) is underway. Missing value reconstruction on the basis of oblique local factorization has been attempted on the Netflix Challenge corpus: my results with 15 extracted axes are encouraging, but a further scaling effort is demanded for a significative RMSE improvement of my initial submission ("Le_Lu" team in the Netflix leaderboard).

7. Contracts and Grants with Industry

7.1. Collaboration with Crédit Agricole S.A.

Participants: Anne Boyer, Armelle Brun, Ilham Esslimani.

The industrial contract between the Crédit Agricole Banking Group (Head Office - Paris) and Kiwi Team has been established for a duration of three years, from November 2006 until the end of 2009. It represents a research contract which aims at suggesting new Collaborative Filtering and social navigation techniques exploiting navigational patterns, adapted to the banking context. The finality of integrating these techniques consists in the optimization of the usage of intranet resources by guiding users towards relevant ones corresponding to their profiles.

7.2. Collaboration with Alcatel Lucent

Participants: Anne Boyer, Armelle Brun, Cedric Bernier.

The contract between Alcatel-Lucent Bell Labs (the "Infrastructure Research and Development" team) and the KIWI team is concretized by a CIFRE grant for a Phd Student. The CIFRE grant has started on the 1st of February 2008 and it will expire the 31st of January 2011. The collaboration concerns the proposition of a model of recommendation for groups, community modeling and evolution analysis.

8. Dissemination

8.1. Scientist Community Animation

8.1.1. Journal and Conference reviewing

- Sylvain Castagnos was a reviewer for CHI2008
- Alain Lelu was a reviewer for the MCM (Mathematical and Computer Modelling) journal
- Armelle Brun was a reviewer for the CAP 2008 conference
- Anne Boyer and Azim Roussanaly were reviewers for the CIDE2008 conference and the TSI journal

8.1.2. Conference organization, Program committees, Editorial boards

- Armelle Brun is a member of the program committee of the CAP 2008 conference
- Anne Boyer is a member of the program committee of the CIDE 2008 conference and the TSI journal
- Azim Roussanaly, Anne Boyer and Armelle Brun are members of the programme committee of the SIIE conference
- Alain Lelu is a member of the program committee of the 9e Journées Internationales d'Analyse Statistique des Données Textuelles (Lyon, 12-14 mars 2008),

8.1.3. PhD and HDR committees

Anne Boyer was a reviewer of the Phd Thesis of Afshin Nikseresht from Institut National Polytechnique de Nantes (October, 2008).

8.1.4. Specialist Committees (commission de spécialistes)

- Azim Roussanaly is member of the specialist committee, section 27, of Nancy2 University
- Alain Lelu is member of the specialist committee, section 71, of Jules Verne (Picardie-Artois) University.

8.1.5. Invited speaker

- Alain Lelu was an invited speaker in the Rencontres Inter-Associations Informatique-Statistique (RIAS'08) in Toulouse
- Anne Boyer was an invited speaker in an INRIA seminar in Dijon, entitled "analyse des usages pour améliorer l'accés aux ressources".
- Anne Boyer was an invited speaker at the autumn school nosstia in Homs (Syria)
- Anne Boyer is a selected speaker of the CIUEN (Colloque International de l'Université á l'Ere du Numérique) conference, in the round table "Elaborer une stratégie d'établissement en s'appuyant sur les politiques nationales, régionales et locales".
- Anne Boyer was an invited speaker of the SaarLorLux conference (e-learning conference of the SaarLorLux chart)

8.1.6. Other responsibilities

- Sylvain Castagnos was member of the lab commitee
- Armelle BRUN was a member of the "ingeneer comipers", the INRIA Lorraine LORIA examination comittee for fixed-term position engineer.
- Anne Boyer was a reviewer for project in the frame of the ANR program

- Anne Boyer is the vice-president of the AUNEGE (Association des Universités pour l'enseignement numérique en économie-gestion) thematic university
- Anne Boyer is the head of the Nutice (Nancy-Université technologie de l'information et de la communication au service de l'enseignement) inter-University service
- Anne Boyer was the chargée de mission TICE of the Nancy2 University
- Anne Boyer is a member of the strategic commitee of the mediteranean virtual university
- Anne Boyer is member of the selection jury of engineers in computer science at Nancy2 University
- Anne Boyer is member of the Administration council of URFIST
- Anne Boyer is member of director board of the LNTLC european project
- Anne Boyer is member of the Administration council of CIRIL
- Anne Boyer and Sylvain Castagnos are creators of the Sailendra SAS start-up (born the 1st january, 2008 and are its scientific council.

8.2. Teaching

- Anne Boyer teaches a datamining course in Master of computer science, 2nd year.
- Azim Roussanaly teaches a course about intelligent assistants in 2nd year of master research in cognitive science
- Azim Roussanaly teaches web services in 2nd year of computer science master
- Alain Lelu teaches the course "statistical methods applied to text mining" in 2nd year of master "information systems", option: technical and scientific information (Université de Marne-la-Vallée)

9. Bibliography

Major publications by the team in recent years

[1] G. BONNIN, A. BRUN, A. BOYER. *Using Skipping for Sequence-Based Collaborative Filtering*, in "IEEE/WIC/ACM International Conference on Web Intelligence, Australie Sydney", University of Technology, Sydney, Australia, 2008, http://hal.inria.fr/inria-00332238/en/.

[2] S. CASTAGNOS, A. BOYER. *Privacy Concerns when Modeling Users in Collaborative Filtering Recommender Systems*, in "Social and Human Elements of Information Security: Emerging Trends and Countermeasures", M. GUPTA, R. SHARMAN (editors), IdeaGroup, Inc., 2008, http://hal.inria.fr/inria-00171806/en/.

[3] S. CASTAGNOS, A. BRUN, A. BOYER. *Probabilistic Reinforcement Rules for Item-Based Recommender Systems*, in "18th European Conference on Artificial Intelligence (ECAI 2008), Grèce Patras", ECCAI (editor), University of Patras, 2008, http://hal.inria.fr/inria-00329560/en/.

[4] A. LELU, M. CADOT. *Graphes des liens et anti-liens statistiquement valides entre les mots d'un corpus textuel*, in "Extraction et gestion de connaissance 2009 (EGC'09), Strasbourg", J.-G. GANASCIA (editor), Pierre Gançarski, 2009, http://hal.inria.fr/inria-00342751/en/.

Year Publications

Doctoral Dissertations and Habilitation Theses

[5] S. CASTAGNOS. *Modélisation de comportements et apprentissage stochastique non supervisé de stratégies d'interactions sociales au sein de systèmes temps réel de recherche et d'accès à l'information*, Ph. D. Thesis, Université Nancy II, 11 2008, http://tel.archives-ouvertes.fr/tel-00341470/en/.

Articles in International Peer-Reviewed Journal

[6] I. ESSLIMANI, A. BRUN, A. BOYER. *Behavioral similarities for collaborative recommendations*, in "Journal of Digital Information Management", 2009-01-01, 11, http://hal.archives-ouvertes.fr/hal-00337730/en/.

International Peer-Reviewed Conference/Proceedings

[7] G. BONNIN, A. BRUN, A. BOYER. *Collaborative Filtering inpired from Language Modeling*, in "International Workshop on Recommender Systems and Personalized Retrieval (RSPR), Tchèque, République Ostrava", 2008, http://hal.inria.fr/inria-00327070/en/.

[8] G. BONNIN, A. BRUN, A. BOYER. *Using Skipping for Sequence-Based Collaborative Filtering*, in "IEEE/WIC/ACM International Conference on Web Intelligence, Australie Sydney", University of Technology, Sydney, Australia, 2008, http://hal.inria.fr/inria-00332238/en/.

[9] G. BONNIN, A. BRUN, A. BOYER. *A Low-Order Markov Model integrating Long-Distance Histories for Collaborative Recommender Systems*, in "International Conference on Intelligent User Interfaces (IUI), États-Unis d'Amérique Sanibel Island", 2009, http://hal.inria.fr/inria-00341537/en/.

[10] M. CADOT, A. LELU. *Massive Pruning for Building an Operational Set of Association Rules: Metarules for Eliminating Conflicting and Redundant Rules.*, in "International Conference on Information, Process, and Knowledge Management - eKnow09, Mexique Cancun", 2008, http://hal.inria.fr/inria-00337067/en/.

[11] S. CASTAGNOS, A. BRUN, A. BOYER. *Probabilistic Association Rules for Item-Based Recommender Systems*, in "4th European Starting AI Researcher Symposium (STAIRS 2008), in conjunction with the 18th European Conference on Artificial Intelligence (ECAI 2008), Grèce Patras", University of Patras, 2008, http://hal.inria.fr/inria-00329559/en/.

[12] S. CASTAGNOS, A. BRUN, A. BOYER. *Probabilistic Reinforcement Rules for Item-Based Recommender Systems*, in "18th European Conference on Artificial Intelligence (ECAI 2008), Grèce Patras", University of Patras, 2008, http://hal.inria.fr/inria-00329560/en/.

[13] Q. T. DINH, H. P. LE, T. M. H. NGUYEN, C. T. NGUYEN, M. ROSSIGNOL, X. L. VU. *Word segmentation of Vietnamese texts: a comparison of approaches*, in "6th international conference on Language Resources and Evaluation - LREC 2008, Maroc Marrakech", ELRA - European Language Resources Association, 2008, http://hal.inria.fr/inria-00334760/en/.

[14] I. ESSLIMANI, A. BRUN, A. BOYER. *Enhancing Collaborative Filtering by frequent usage patterns*, in "First International Conference on the Applications of Digital Information and Web Technologies - ICADIWT 2008, Tchèque, République Ostrava", 2008-08-04, p. 180-185, http://hal.archives-ouvertes.fr/hal-00337728/en/.

[15] H. P. LE, T. V. HO. *A Maximum Entropy Approach to Sentence Boundary Detection of Vietnamese Texts*, in "IEEE International Conference on Research, Innovation and Vision for the Future - RIVF 2008, Viet Nam Ho Chi Minh City", 2008, http://hal.inria.fr/inria-00334762/en/.

[16] H. P. LE, T. MINH HUYEN NGUYEN, A. ROUSSANALY. *A Metagrammar for Vietnamese LTAG*, in "9th International Workshop on Tree Adjoining Grammars and Related Formalisms - TAG+9 2008, Allemagne Tubingen", 2008, http://hal.inria.fr/inria-00334752/en/.

[17] H. P. LE, T. M. H. NGUYEN, A. ROUSSANALY, T. V. HO. *A Hybrid Approach to Word Segmentation of Vietnamese Texts*, in "2nd International Conference on Language and Automata Theory and Applications - LATA 2008 Language and Automata Theory and Applications Lecture Notes in Computer Science, Espagne Tarragona", vol. 5196, Springer Berlin / Heidelberg, 2008, p. 240-249, http://hal.inria.fr/inria-00334761/en/.

National Peer-Reviewed Conference/Proceedings

[18] A. LELU, M. CADOT. *Graphes des liens et anti-liens statistiquement valides entre les mots d'un corpus textuel*, in "Extraction et gestion de connaissance 2009 (EGC'09), France Strasbourg", J.-G. GANASCIA (editor), Pierre Gançarski, 2009, http://hal.inria.fr/inria-00342751/en/.

Workshops without Proceedings

[19] M. ZITT, E. BASSECOULARD, A. LELU. *Hybrid maps of scientific fields: an application to nanosciences*, in "10th International Conference on Science and Technology Indicators (S&TI 2008), Autriche Vienne", Anthony Van Raan, 2008, http://hal.inria.fr/inria-00334304/en/.

Scientific Books (or Scientific Book chapters)

[20] A. BOYER. *Analyse des usages pour améliorer l'accès aux ressources*, in "Métadonnées : mutations et perspectives : séminaire INRIA, 29 septembre - 3 octobre 2008, Dijon Sciences et techniques de l'information", L. CALDERAN, B. HIDOINE, J. MILLET (editors), ADBS éditions, 2008, p. 89-112, http://hal.inria.fr/inria-00341479/en/.

[21] S. CASTAGNOS, A. BOYER. *Privacy Concerns when Modeling Users in Collaborative Filtering Recommender Systems*, in "Social and Human Elements of Information Security: Emerging Trends and Countermeasures", M. GUPTA, R. SHARMAN (editors), IdeaGroup, Inc., 2008, http://hal.inria.fr/inria-00171806/en/.

[22] A. LELU. *Visualiser les textes et les mots : approches numériques, approches par les graphes*, in "Information et visualisation, contribution à l'ergonomie visuelle", S. CHAUVIN (editor), CEPADUES, Toulouse, 2008, http://hal.inria.fr/inria-00334267/en/.

Other Publications

[23] A. LELU. *La méthode de classification non-supervisée K-means axiales*, 2008, http://hal.inria.fr/inria-00333865/en/.

References in notes

[24] A. BANERJEE, I. S. DHILLON, J. GHOSH, S. SRA. *Clustering on the Unit Hypersphere using von Mises-Fisher Distributions*, in "Journal of Machine Learning Research", 2005.

[25] M. CADOT. *A Randomization Test for extracting Robust Association Rules*, in "3rd world conference on Computational Statistics & Data Analysis - CSDA 2005, Limassol Chypre", 2005, http://hal.inria.fr/inria-00337069/en/.

[26] S. CHAN, J. GOODMAN. *An empirical study of smoothing techniques for language modeling*, Technical report, Harvard University, august 1998.

[27] A. LELU, M. CADOT, P. CUXAC. *Document stream clustering : an optimal and fine-grained incremental approach*, in "COLLNET'06, France Nancy", 2007, http://hal.inria.fr/inria-00333876/en/.

[28] E. STREHL, J. GHOSH, R. MOONEY. *Impact of similarity measures on web-page clustering*, in "In Workshop on Artificial Intelligence for Web Search (AAAI 2000", AAAI, 2000, p. 58–64.

[29] M. ZITT, E. BASSECOULARD, A. LELU. *Mapping nanosciences by citation flows: a preliminary analysis*, in "Scientometrics", 2007, http://hal.inria.fr/inria-00334331/en/.