Conference Chair - Welcome Message

Dear AIDD 2015 Attendees,

It is my great pleasure to express my warmest welcome to all participants to AIDD'2015, which is organized by the laboratory of research in artificial intelligence (LRIA) at USTHB. The aim of this event is to help Ph. D student to keep abreast of the latest developments in the area of artificial intelligence and exchange ideas about their research work. This year, AIDD'2015 will focus on High Performance Computing (HPC) for Artificial Intelligence and Data Mining, seeing that in recent years HPC is involved in many fields and is perfectly aligned with artificial intelligence since both technologies appeared to effectively solve the problems of everyday life. Thanks to the supercomputers that include thousands of processors to treat tremendous volumes of data. In absence of such sophisticated device, the advanced technology that we are knowing nowadays, would not have emerged. Artificial intelligence also contributed widely in solving complex and hard problems. One important idea to investigate is to combine these two advanced fields to yield more progress in sciences and technologies. I am especially grateful to the keynote speakers who accepted to give an invited lecture on such up-to-date subject and to the authors for their valuable papers.

As usual, let me express my profound gratitude to the organizing committee for having used their best endeavors to the success of the event. I also thank the program committee and the reviewers for their contribution to evaluate the submissions. Thanks also to USTHB leaders who helped in terms of logistics.

Finally, thank you for joining AIDD'2015, we sincerely hope you will enjoy your stay and wish you a fruitful scientific meeting.

Prof Habiba DRIAS
AIDD'2015
Conference Chair

Program Committee Chair - Welcome Message

I have a great pleasure to welcome you to the 4th edition of Artificial Intelligence Doctoral Days (AIDD'2015). This scientific conference is organized by the Artificial Intelligence Research Laboratory (LRIA) at the University of Sciences and Technology Houari Boumediene.

AIDD'2015 is a means of transferring knowledge and intense scientific exchange between nationals and internationals researchers in the field of artificial intelligence. This scientific event gives to PhD students the opportunity to present and promote their research work and enrich their scientific knowledge by attending conferences given by the invited speakers on various topics.

The program of AIDD'15 consists of twelve oral presentations organized into four sessions and six plenary talks given by top researchers, each of them having an excellent track record in his or her own field of research. I thank them ail sincerely for accepting our invitation and wish them a pleasant stay in Algiers:

- Mr Mohamed Hannache, Head of Innovation Division, Ministry of Industry and Mines.
- Prof Lakhdar Sais, CRIL - CNRS, "Université d'Artois", Lens, France.
- Dr Boualem Bouzid, SNIRM, "Faculty of Physics, USTHB.
- Dr Ahcene Latreche, Director Strategic Projects & Global Offers – Bull France, Groupe Atos.
- Dr Rachid Gherbi "University Paris XI Orsay" France.
- Mrs Karima Ibelaidene, Researcher & Management Quality Supervisor, Sonatrach, Algeria.

I also thank the authors who entrusted us with their nice and valuable contributions. My sincere and deep gratitude goes also to the program committee members for the rigorous review of the manuscripts. I thank also all the participants of the conference and the organizing committee for their valuable support and contribution in the preparation of the AIDD'2015 conference.

We welcome all to have very rewarding scientific days of interaction and very pleasant moments in Algiers. On behalf of all AIDD'2015 committee, we wish you very productive days with fruitful discussions.

Dr Samir KECHID
AIDD’2015
Program Committee Chair

Organizing Committee

Organizing Committee Chair: Ms. Hadia Mosteghanemi
Members:

Ms. Aicha Boutorh
Ms. Amel Ourahmoune
Ms. Asma Bellili
Mr. Bachir Bahamida
Ms. Hireche Célia
Mr. Ibrahim Chegrane
Ms. Marwa Djeffal
Mr. Mohamed Amine Chemchem
Mr. Nadjib Fodil-Cherif
Ms. Neyla Benhamouda
Ms. Raouia El Nagger
Mr. Yassine Drias
Mr. Zakaria Saoud

Program Committee

Program Committee Chair: Dr. Samir Kechid
Members:

Prof. Ahmed Guessoum
Prof. Dalila Boughaci
Prof. Habiba Drias
Prof. Malika Boukala-Ioualalen
Prof. Slimane Larabi
Prof. Thouraya Bouabana Tebibel
Dr. Cherif Boukala
Dr. Djamila Dahmani
Dr. Feryel Souami
Dr. Hamid Azzoune
Dr. Hamid Necir
Dr. Lamia Berkani
Dr. Nacéra Bensaou
Dr. Nacéra Laiche
Dr. Nadjet Kamel
Dr. Nadia Baha
Dr. Nassim Zellal
Dr. Sadjia Benkhider
Dr. Saliha Aouat

LRIA AIDD'2015 SCHEDULE

Conference room of "the Science Home"

Tuesday, March 17, 2015

08:30	**Welcome**
09:00	**Opening Ceremony**
09 :30	**Plenary session 1 :** Chair Prof Habiba Drias & Prof Malika Boukala-loualalene Invited Speaker **: Mr Mohamed Hannache, «** Chef de Division de l'Innovation, Ministère de l'Industrie et des Mines » ***« Les TIC, levier numérique de l'innovation »***
10 :15	**Plenary session 2 :** Chair Prof Thouraya Tebibel & Dr Faiza Khellaf Invited Speaker : **Prof Lakhdar Sais,** CRIL – CNRS, « Université d'Artois », Lens, France ***« Building Bridges between Data Mining and Artificial Intelligence »***
11:15 Coffee Break	
Session 1: Chair Prof Slimane Larabi & Prof Amar Aissani	
11:45	Meriem Khelifa, *A hybrid Method for the Traveling Tournaments Problem*
12:15	Abdelhak Bousbaci, *Parallel Sampling-PSO-Multi-Core-K-Means Using Mapreduce*
13:00 Lunch	
14:00	**Plenary session 3 :** Chair Prof Zaia Alimazighi & Prof Ahmed Guessoum Invited Speaker : **Dr Boualem Bouzid**, SNIRM, «Faculté de Physique, USTHB» ***« L'usage du HPC en médecine nucléaire »***
15:00 Coffee Break	
Session 2: Chair Prof Dalila Boughaci & Dr Samir Kechid	
15:30	Karimal belaidene*, Reformulation's Optimization of Unleaded Premium Gasoline with Ethanol*
16:00	Yassine Drias, *Hybrid ACO and Tabu Search for Web Information Foraging*
16:30	Zakaria Saoud, Improving Source *Selection Process using Social Profile*
17:00	Hamid Benachour, *Contextual Source Selection for Federated Search in Mobile Environment*

LRIA AIDD'2015 SCHEDULE

Conference room of "the Science Home"

Wednesday, March 18, 2015

09:00	**Plenary session 4 :** Chair Prof Aicha Mokhtari & Dr Nadjet Kamel Invited Speaker **: Dr Ahcene Latreche,** Director Strategic Projects & Global Offers – Bull France, Groupe Atos ***« Le Calcul Intensif : la fondation des économies de l'innovation »***
10 :00	**Plenary session 5 :** Chair Prof Djamel Zegour & Dr Feryel Souami Invited Speaker : **Dr Rachid Gherbi** « Université Paris XI Orsay », France ***« La Réalité Virtuelle au Service du Visual Mining »***
11:00 Coffee Break	
Session 3: Chair Dr Nadia Baha & Dr Saliha Aouat	
11:30	Asma Bellili, *Image Segmentation by Image Analogies*
12:00	Izem Hamouchenne, *Texture Analysis and Matching*
12:30	Amel Berrachedi, *Evaluation of the packet loss in WSN using Deterministic Stochastic Petri Nets*
13:00 Lunch	
14:00	**Plenary session 6 :** Chair Prof AbdelKader Belkhir & Prof Walid Hidouci Invited Speaker : **Mrs Karimal belaidene**, Researcher & Management Quality Supervisor , Sonatrach, Algeria ***« How can Fuel Productions take the green route »***
15:00 Coffee Break	
Session 4: Chair Dr Hamid Azzoune & Dr Nacéra Bensaou	
15:30	Imene Messaoudi, *Bat algorithm for overlapping community detection*
16:00	Abderahmane Khiat, *Instance Matching Tools for Linked Data: A Comparative Study*
16:30	KamelEddine Heraguemi, *Modified Bat Algorithm for Mining Association Rule*
17:00 Closing Ceremony	

Tables of Content

Plenary Talks' Abstracts

Presented Papers

Plenary Talk 1 : *Les TIC, levier numérique de l'innovation*

ABSTRACT

L'innovation est désormais unanimement considérée comme l'un des éléments clés de la stratégie et des politiques des firmes et même des États. Autrement dit, les avantages comparés des nations et des entreprises, notamment industrielles, ne se limitent plus aux avantages constitués par les ressources naturelles ou la main-d'œuvre, mais sont de plus en plus portés par l'innovation. Mais si tout le monde admet cette nouvelle réalité, ce qui semble encore moins perçu et admis, c'est que la maîtrise de l'innovation est elle-même de plus en plus dépendante de la maîtrise des technologies de l'information et de la communication (TIC).

C'est pourquoi il est aujourd'hui impératif, aussi bien pour les chercheurs que pour les industriels, de recourir massivement aux TIC - et plus encore aux possibilités offertes par le calcul intensif - pour élargir les limites de leurs connaissances ou pour développer leurs produits innovants. Ce recours massif aux TIC sera d'autant plus salutaire aux chercheurs et aux industriels qu'il leur permettra, à terme, de réduire les coûts de leurs dépenses ou de leurs investissements tout en augmentant la fiabilité de leurs approches. Le calcul intensif peut s'avérer un puissant et précieux support à la promotion de l'innovation industrielle, celle-ci constituant, précisément, l'un des axes majeurs de la nouvelle stratégie industrielle amorcée par l'Algérie.

Mr Mohamed Hannache

Biography

Économiste de formation, Mohammed Hannache occupe actuellement (2012-2015) la fonction de Chef de Division de l'Innovation au ministère de l'Industrie et des Mines (MIM). Il a auparavant occupé les fonctions de Chef de Division de la Promotion de l'Utilisation des TIC (2010-2012) au ministère de l'Industrie et de la Promotion de l'Investissement (MIPI) et de Directeur d'études (2002-2010) auprès du même ministère. Il a contribué à divers travaux portant sur le rôle des TIC comme levier numérique de l'innovation et de la promotion de la compétitivité industrielle en Algérie. En 2008, il a représenté le MIPI aux travaux de la e-Commission, structure chargée de l'élaboration du document e-Algérie.

De 1986 à 2002, il a occupé les postes de Responsable d'études principal, puis de Directeur de l'Information et de la Communication à l'Agence Nationale pour l'Aménagement du Territoire (ANAT). Au sein de l'ANAT, il a contribué à l'élaboration de diverses études, dont le Schéma National d'Aménagement du Territoire (SNAT) ; la Carte scolaire de la wilaya d'Alger ; la Carte de la formation professionnelles de la Wilaya d'Alger…

En parallèle, Mohammed Hannache intervient en tant qu'enseignant associé auprès de l'Institut National de la Poste et des TIC (INPTIC). Modules enseignés : Management de l'Innovation, Marché et Pratiques de l'Industrie de l'Information et Management de la Qualité.

Plenary Talk 2 : *Building Bridges between Data Mining and Artificial Intelligence*

ABSTRACT

In this talk, we overview our contribution to data mining and more generally to the cross-fertilization between data mining, constraint programming and propositional satisfiability (http://www.cril.univ-artois.fr/decMining/). We will focus on two contributions. First, we show how propositional satisfiability can be used to model and solve problems in data mining. As an illustration, we present a SAT-based declarative approach for mining sequences. Secondly, we discuss how symmetries widely investigated in Constraint Programming (CP) and Propositional Satisfiability (SAT) can be extended to deal with data mining problems.

Prof Lakhdar Sais

Biography

Prof Lakhdar Sais obtained an engineering degree in computer science in 1988 from the National Institute on Computer Science ("Université de Tizi-Ouzou", Algeria), a Ph.D ("Doctorat") in 1993 from the "Université de Provence" (Marseille) and an "Habilitation à Diriger des Recherches" from the "Universitéd'Artois" in 2000. In 1994, he joined the "IUT de Lens" as a lecturer ("Maitre de conférences") at the beginning of the creation of the CRIL research center ("Centre de RechercheenInformatique de Lens"). Before his current position as a professor at CRIL-CNRS "Universitéd'Artois", he spent one year as a professor at IRIT «Université Paul Sabatier » (Toulouse, France). He spent two years as a researcher at INRIA Lille and CNRS. He is the founding-member and the leader (from 2002 – 2013) of the inference and decision process group at CRIL - CNRS. He is currently the Delegate director of the CRIL laboratory. His research focuses on search and representation problems in Artificial Intelligence. He is especially interested in propositional satisfiability, quantified boolean formula, constraint satisfaction and optimisation problems, knowledge representation and reasoning, data mining. For further details on his research activities visit the web site: http://www.cril.fr/~sais.

.

Plenary Talk 3 : *L'usage du HPC en médecine nucléaire*

ABSTRACT

L'imagerie nucléaire est une imagerie par émission utilisée en médecine nucléaire. Elle consiste à déterminer la distribution dans l'organisme d'une substance radioactive administrée au patient, appelée radiotraceur, en détectant le rayonnement qu'elle émet au moyen d'un dispositif de détection adapté à un rayonnement externe. Deux modalités d'imagerie sont couramment employées en imagerie nucléaire: la tomographie d'émission monophotonique (SPECT : Single Photon Emission ComputedTomography), pour laquelle le radiotraceur émet des photons gamma détectés grâce à une gamma caméra, et la tomographie à émission de positons (PET : Positron Emission Tomography), pour laquelle le radiotraceur émet des positons et où l'on détecte les deux photons gamma émis en coïncidence après annihilation du positon avec un électron. Les données acquises par le détecteur sont reconstruites à l'aide d'algorithmes de reconstruction afin de fournir une estimation de la distribution tridimensionnelle du radiotraceur dans l'organisme. L'imagerie nucléaire permet d'avoir accès à des informations sur le fonctionnement des organes et d'étudier des processus physiologiques et métaboliques. La fiabilité de la quantification des images obtenues est affectée à la fois par les limites des performances des détecteurs (résolution spatiale et en énergie, sensibilité, …), par les effets physiques tels que l'atténuation, la diffusion et l'effet de volume partiel (qui perturbent la formation des images), par des effets physiologiques (mouvements respiratoire et cardiaque) et par des effets liés à la reconstruction tomographique. Ces effets doivent donc être corrigés par des méthodes de correction spécifiques afin d'extraire des paramètres quantitatifs fiables.

Dans ce cadre, les simulations Monte-Carlo représentent un outil puissant et efficace d'aide à l'optimisation des composants des détecteurs (collimateur, géométrie, …), à la conception de nouveaux détecteurs, au développement et à l'évaluation des algorithmes de reconstruction et de méthodes de corrections des effets physiques. De nombreux codes de simulation Monte-Carlo sont actuellement disponibles. Cependant, leur inconvénient principal est le temps de calcul nécessaire à leur exécution. Actuellement, l'arrivée d'ordinateurs de plus en plus puissants et les possibilités de partage de ces calculs sur les clusters et/ou les grilles informatiques permettent de réduire considérablement ces temps de calcul pour une utilisation efficace en recherche dans le domaine de la médecine nucléaire.

Au cours de ces journées, nous allons vous présenter un retour d'expérience de notre utilisation du HPC (Cluster SNIRM.PhysUSTHB.dz, Grille Dz e-Science et Grille VIP /GateLab) pour réaliser différentes études dans le domaine de la médecine nucléaire.

.

Dr Boualem Bouzid

Biography

Dr Boualem Bouzid est Maître de Conférences A à la Faculté de Physique de l'USTHB, il a une longue expérience dans l'enseignement et la recherche scientifique au sein respectivement de la Faculté de Physique et du laboratoire SNIRM. Il est titulaire d'un Doctorat d'Etat en Physique Nucléaire. Il a contribué à divers thèmes de la physique nucléaire (fission nucléaire, straggling-pouvoir d'arrêt, astrophysique nucléaire et physique médicale). Depuis cinq ans, Il s'est consacré à la physique médicale et particulièrement à la modélisation des systèmes nucléaires d'imagerie médicale utilisant les techniques de Monte Carlo, de dosimétrie, de correction d'image, de reconstruction et de quantification en tomographie SPECT, PET et CT. Ces contributions scientifiques ont fait l'objet de plusieurs publications et communications. Actuellement, il s'intéresse comme utilisateur potentiel au techniques HPC.

Plenary Talk 4 : *Le Calcul Intensif : la fondation des économies de l'innovation*

ABSTRACT

Le Calcul Intensif (ou HPC : High Performance Computing) est au cœur d'une révolution culturelle et d'un changement de paradigme dans le monde de la recherche et de l'industrie, où le duo *théorie – expérimentation* est remplacé par un trio *théorie – simulation/modélisation-expérimentation.* Par sa caractéristique d'être transverse à tous les secteurs de l'économie, le Calcul Intensif est devenu indissociable de toute perspective de croissance, d'innovation, de création d'emplois et de compétitivité.

Le Calcul Intensif est un accélérateur d'innovation permettant à un pays et aux industries d'exploiter leurs atouts et leurs moyens pour prendre part à la compétitivité économique des nations.

Le Calcul intensif est un écosystème pluridisciplinaire, regroupant les sciences du calcul scientifique, les logiciels de la simulation et de la modélisation, les scientifiques presque de toutes les disciplines et les infrastructures à très haute performance. Cet écosystème est engagé dans une spirale de la course à la puissance et à la complexité. Plus grands sont les problèmes à résoudre, plus grande est la puissance de calcul nécessaire aux besoins. Plus grande est la puissance de calcul disponible, plus grands sont les problèmes qu'on cherche à résoudre. Et cette spirale à la puissance et à la complexité pose des défis majeurs aux acteurs du calcul intensif que nous sommes.

Dr Ahcene Latreche

Biography

Ahcene est diplômé d'un Doctorat (PhD) et d'un Master Exécutif en Management Général International de l'ESSEC Business School Paris.

Ahcène possède plus de 20 ans d'expérience à différents postes de responsabilité en environnement international dans les missions de conseil, de management de projets stratégiques, de management de l'avant-vente et de programmes marketing stratégie, dans les domaines de la sécurité et des logiciels de management des infrastructures informatiques des entreprises et de l'industrie des télécommunications.

Au sein du groupe Atos/Bull, Ahcène est actuellement en charge de contribuer au développement du leadership du groupe Atos/Bull dans les secteurs stratégiques du Big Data, du HPC et de la Cybersecurité.

Plenary Talk 5 : *La Réalité Virtuelle au Service du Visual Mining*

Dr Rachid Gherbi

Biography

Rachid Gherbi est Enseignant-Chercheur en Informatique à l'Université Paris-Sud XI Orsay, France, depuis 1988. Il est né en 1963 à Alger et a obtenu son diplôme d'ingénieur d'état en informatique à l'USTHB (Université de Science et de technologie Houari Boumediene) en 1987. Reçu au concours national de bourses de post-graduation, il est parti effectuer un DEA (Master recherche) en Fondements des Systèmes Informatiques en 1988 et un Doctorat (Ph. D) en Vision par machine en 1992 à l'Université Paris-Sud. Il a obtenu son HDR (Habilitation à Diriger les recherches) en 2001 dans la même université sur les thématiques de la représentation et du traitement de données complexes en IHM. Il a co-fondé et dirigé plusieurs groupes de recherche en communication homme-machine dans les laboratoires LIMSI-CNRS à Orsay et Genopole à Evry. Il a a encadré une dizaine de doctorants et a publié une centaine d'article internationaux. Il a géré plusieurs projets financés au niveau national et européen, en partenariat avec des équipes académiques et des industriels, ainsi que l'organisation de conférences et de journées scientifiques. Il a participé activement dans les différentes instances universitaires de formation et de recherche. Il a été Professeur invité à Concordia Université (Canada) en 2005. Plusieurs formations universitaires en Licence et en Master ont été créées et conçues par Rachid Gherbi tout au long des ces vingt dernières années. Ses thématiques de recherche vont de la vision par machine à la réalité virtuelle et augmentée dans le contexte de l'interaction homme-machine.

ABSTRACT

Le Visual Mining (VM) peut être défini comme le processus cognitif qui intègre l'humain dans l'analyse des informations en utilisant des systèmes de visualisation interactifs. La Réalité Virtuelle (RV) est une simulation informatique interactive immersive, visuelle, sonore et/ou haptique, d'environnements réels ou imaginaires, dont la finalité est de permettre à une personne (ou à plusieurs) une activité sensori-motrice et cognitive dans un monde artificiel, et qui peut être associé au réel. Par ailleurs, les interfaces humain-machine (IHM) ont vu dans les deux dernières décennies un développement majeur de nouveaux moyens/dispositifs de communication.

Je vais aborder dans cette conférence ces trois sujets, poser les problématiques et parcourir les solutions, à la fois matérielles et logicielles, de ces sujets. Il s'agit en premier lieu de montrer que l'immersion des utilisateurs dans de tels environnements virtuels exige une réflexion adaptée et une étude de nouveaux paradigmes d'interaction humain-environnement. Je présenterai au cours de l'exposé les problématiques et les pistes d'étude et de recherche au croisement de ces différents domaines. Plusieurs champs applicatifs seront présentés pour illustrer tout l'intérêt de la combinaison de la Réalité Virtuelle et du Visual Mining.

Plenary Talk 6 : *How can Fuel Productions take the green route*

ABSTRACT

As part of the improvement of Algerian's gasoline quality, this study aims at reformulating unleaded premium gasoline which, not only preserves the environment but also respects the new international specifications by introduction of ethanol.

For this purpose, we have prepared gasoline with blending components used in the premium gasoline formulation by suppressing lead. On the light of this data, we have established a correlation which enables to predict the octane number value.

In addition, an optimization of different experiments resulted in determining the proportions that should be mixed so that the octane number would be maximized. They also gave us an idea about the catalytic reforming unit's effectiveness.

Mrs Karima Ibelaidene

Biography

Karima IBELAIDENE, KIROUANI, undertook her initial degrees in industrial Chemistry from the National Institute of Hydrocarbon and Chemistry of Boumerdes (Algeria). She worked as an Engineer Analysis in Research and Development Center of SONATRACH (1989-2008) before joining, in 2008, the Quality Management Department.

She completed a Post Graduation in Refining & Petrochemistry and Magister from Polytechnic School in 2005 and prepares currently a Doctorate at the University of Blida.

She has also dealt with several research themes relating to petroleum activity, supervised numerous Master and engineering thesis projects,.

he is an internal auditor and technical appraiser- ALGERAC Expert and presented many communications in different conferences.

Presented Papers

A hybrid Method for the Traveling Tournaments Problem

Meriem Khelifa and Dalila Boughaci

LRIA, USTHB,
USTHB, BP 32 El Alia, Bab Ezzouar Algiers, Algeria

khelifa.merieme.lmd@gmail.com, dboughaci@usthb.dz

Abstract. In this work, we propose an original search method for the Traveling Tournaments Problem (TTP) which is a fundamental timetabling problem in sports scheduling. The proposed method consists of three phases. In the first phase, we used the theory graph modeling to create an initial configuration that satisfies only the Double Round Robin Tournament (DRRT) constraint. In the second phase, we used Iterated Local Search (ILS)method that takes the initial configuration of phase one as an input, and tries to find feasible configurations with regards to issues of home/away patterns. The third phase is a stochastic local search (SLS) that manipulates feasible configurations and where the main objective is to find an optimal solution that minimizes the distance traveled by all teams. The method is evaluated on Benchmarks and compared with other techniques for TTP. The computational results are promising and show the effectiveness of the proposed approach.

Keywords: sport scheduling, graph theory, Traveling tournament problem, Iterated Local Search (ILS), Local Search(SLS), CSP.

1 Introduction

Sport schedule is an attractive research area that has received considerable attention in recent years. In this paper, we are interested in the traveling tournament problem (TTP) which is an interesting problem in both sport scheduling and combinatorial optimization. TTP is the problem of scheduling a double round-robin tournament, while satisfying a set of related constraints and minimizing the total distance traveled by all teams [4].

The problem can be stated as follows: let us consider n teams (n even and positive), a double round robin tournament is a set of games in which every team plays every other team exactly once at home and once away. A double round robin tournament has 2*(n-1) slots.

The teams begin in their home city and must return there after the tournament. The TTP is the problem of finding a feasible schedule that minimizes the distance traveled by all teams, and satisfies the following constraints:

1. **Double Round Robin Tournament constraint** (DRRT): each team plays with every other team exactly twice, once in his Home and once in the Home of his opponent.
2. **AtMost constraint:** each team must play no more than *U* and no less than *L* consecutive games at home or away. In general, *L* is set to 1 and *U* to 3.
3. **NoRepeat constraint:** A game (t_i, t_j) can never be followed in the next round by the game (t_j, t_i).

The TTP inputs are: the number n of teams and the Dis distance matrix. The output will be a Double Round Robin Tournament on the n teams respected the three constraints AtMost, NoRepeat and DRRT, and the total distance traveled by the team is minimized. Table 1 gives an example of a schedule for n equals to 4. The sign (-) means that the team plays away.

Table 1: Example of Double Round Robin Tournament with n = 4 Teams

T \ R	0	1	2	3	4	5
1	4	-3	-4	-2	3	2
2	-3	-4	3	1	4	-1
3	2	1	-2	4	-1	-4
4	-1	2	1	-3	-2	3

For example the timetabling of team t1 is as follows: t1 plays against the teams: t4 at home, t3 at away, t4 at away, t2 away, t3 at home, and t2 at home successively. The traveled distance by the t1 is equals to the sum of: dis13 + dis34+ dis42+ dis21.

The traveling tournament problem is very difficult to solve. Several methods have been studied for the TTP. Among these methods, we can find the following works:[7] used a "Branch and price" approach for solving the TTP. [5] worked on the problem TTPPV (Traveling Tournament Problem with Predefined Venues), they proposed an Simulated annealing method. [12] proposed an hybrid Branch-and-bound method for TTP. [8] proposed a tabu search approach for TTP with several neighborhood structures[9] .proposed an evolutionary approach for TTP. [2]provided a simulated algorithm that explores feasible and infeasible schedules using several structures of neighborhoods and complex movements.

The current work proposes a novel search method for the TTP. To the best of our knowledge the proposed heuristic approach is original. The originality of our idea is to combine between optimization, graph theory and constraint programming. More precisely, we propose a three-phase search method that starts from an initial solution verifying the DRRT constraint created by using the graph theory modeling. Then this configuration is passed to the second phase that uses a Iterated Local Search(ILS). The ILS main objective is to find a configuration that verifies the three constraints. Here, we used the CSP (constraint satisfaction problem) to model the problem and a cost function to penalize the unfeasible configurations. Finally, the third phase is a stochastic local search (SLS) that starts from the configurations generated by ILS and tries to find the optimal solution that minimizes the total distance traveled by all the teams. SLS used the cost function to verify the feasibility of configurations. It uses also an objective function to measure the quality of configurations. The quality of a

configuration (a schedule) is measured by the total distance traveled by teams.

The rest of this paper is organized as follows: The second section presents in detail the proposed method for the TTP. Some numerical results are given in the third section. Finally, the fourth section concludes the work.

2 Proposed Approach

The proposed search method for TTP consists of three phases. In the first phase, we created an initial configuration satisfying the DRRT constraint. In the second phase, we applied a Iterated Local Search approach to generate feasible configurations. The third phase is the stochastic local search. The different phases are detailed in following subsections.

A. Phase 1: The Initial Configuration

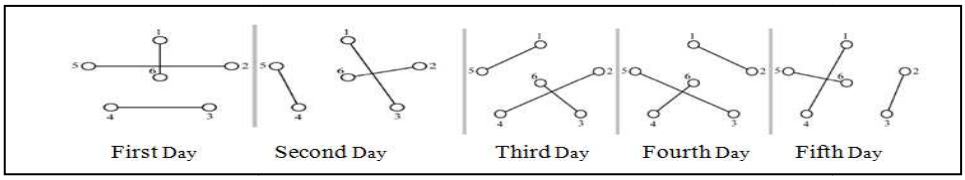
To obtain an initial configuration verifying the DRRT constraint, we have used the theory graph modeling (Werra, 1988).

We have n*(n-1) / 2 games in round n-1 when n is even. We numbered the vertices of the graph from 1 to n where n is the number of teams. We put the top n in the center and the other vertices in a circle around the top n.

• The first day, we organized a game between Team 1 and *n*, Team 2 and *n-1*, Team 3 and *n- 2*, and so on up to the game between *n/2* and *n/2 +1*.

• The following day, we reproduce what happened the previous day, making simply a rotation of the coupling in the direction of clockwise.

Figure 1 depicted the procedure for building a set of $n = 6$ teams.



We obtain the schedule which is a Single Round Robin (SRR)

Round1	(t_1, t_6)	(t_5, t_2)	(t_4, t_3)
Round2	(t_5, t_4)	(t_1, t_3)	(t_6, t_2)
Round3	(t_1, t_5)	(t_2, t_4)	(t_6, t_3)
Round4	(t_1, t_2)	(t_5, t_3)	(t_6, t_4)
Round5	(t_5, t_6)	(t_1, t_4)	(t_2, t_3)

The double round robin is obtained by adding the mirror of SRR. The obtained DRRT is as follows:

Round1	(t_1, t_6)	(t_5, t_2)	(t_4, t_3)
Round2	(t_5, t_4)	(t_1, t_3)	(t_6, t_2)

Round3	(t_1, t_5)	(t_2, t_4)	(t_6, t_3)
Round4	(t_1, t_2)	(t_5, t_3)	(t_6, t_4)
Round5	(t_5, t_6)	(t_1, t_4)	(t_2, t_3)
Round6	(t_6, t_1)	(t_2, t_5)	(t_3, t_4)
Round7	(t_4, t_5)	(t_3, t_1)	(t_2, t_6)
Round8	(t_5, t_1)	(t_4, t_2)	(t_3, t_6)
Round9	(t_2, t_1)	(t_3, t_5)	(t_4, t_6)
Round10	(t_6, t_5)	(t_4, t_1)	(t_3, t_2)

In the initial phase, we have obtained a schedule that respects the double round robin tournament. This configuration will be used in the second phase as an input solution.

B. Phase 2: Iterated local search (ILS)

.In our work, we have used (ILS) in the second phase of our method. After having created a schedule satisfying the DRRT constraint, we call the second phase in which we considered the three constraints DRRT, AtMost and NoRepeat. We used a iterated local search to search for configurations satisfying the three constraints. We used the constraint programming problem (CSP) to modeling the problematic and a cost function to penalize configurations that violate the considered constraints.

B.1 The CSP Modeling

We used the following notations to formulate the problem as a constraint satisfaction problem (CSP):

- R : is a set of rounds $|R| = (n-1)*2$
- T is the set of n teams,
- t_i: team number i, $t_i \in T$, $1 \leq i \leq |T|$
- $x(t_i, t_j)$ *is* the planning game between the teams t_i and t_j in Home of t_i. The values of this variable are of the form $(R_{i,j})$ indicating the round planning game

$R_{i,j} \in R$, $0 \leq R_{i,j} \leq |R|$

The set of the variables are:

$X = \{(x(t_i, t_j),\ 1 \leq i \leq |T|,\ 1 \leq j \leq |T|,\ i \neq j\ \}$

All the domains are equal:

$D = \{R_{i,j}\ \ 0 \leq R_{i,j} \leq |R|\ \}$: $\forall x \in X$, $D_x = D$, where D is the domain of the variable x.

The set of constraints is:

- **The uniqueness of whole teams in all rounds**

For each team $t_{i,}$ $t_i \in T$, $1 \leq i \leq |T|$

Round $(t_j) \longleftrightarrow$

$((x(t_i, t_j) \neq x(t_a, t_j)) \wedge (x(t_i, t_j) \neq x(t_j, t_a)) \wedge (x(t_j, t_i) \neq x(t_a, t_j)) \wedge$

$(x(t_{j,}\ t_i) \neq x(t_j, t_a))$, $\forall (i, a) \in [1, |T|^2\ i$
$i \neq j,\ a \neq j, i \neq a,$

- **The constraint "NoRepeat"**

For each team : $<t_i, t_j>$, $(t_i, t_j) \in T \times T$, $t_i \neq t_j$
NoRepeat ($<t_i, t_j>$) :
(if $(x(t_i, t_j) > 0 \wedge (x(t_j, t_i) < (x(t_i, t_j)) \rightarrow x(t_j, t_i) \neq (x(t_i, t_j)\text{-}1) \wedge$

(if $(x(t_i, t_j) \leq |R|\text{-}1 \wedge (x(t_j, t_i) > (x(t_i, t_j)) \rightarrow x(t_j, t_i) \neq (x(t_i, t_j)+1)$

- **The constraint "AtMost" :**

For each team $t_{i,}$ $t_i \in T$, $1 \leq i \leq |T|$
AtMost:(t_i):

$$\left\{ \begin{array}{l} \left(\begin{array}{l} \text{if } ((x(t_i, t_{a1}) \wedge (x(t_i, t_{a2}) = x(t_i, t_{a1})+1) \wedge (x \\ (t_i, t_{a3}) = x(t_i, t_{a1})+2 \wedge (x(t_j, t_{a4}) = x(t_i, \end{array} \right) \\ \vee \\ \left(\begin{array}{l} \text{if } ((x(t_{a1}, t_i) \wedge (x(t_{a2}, t_i) = x(t_{a1}, t_i,)+1) \wedge (x \\ (t_{a3}, t_i,) = x(t_{a1}, t_i,)+2 \wedge (x(t_{a4}, t_j) = x(t_i, \\ t_{a1})+3)) \end{array} \right) \end{array} \right.$$

$\rightarrow (i \neq j)$, $\forall (a_1, a_2, a_3, a_4, i) \in T \times T \times T \times T \times T$
$a_i \neq a_j$, $i \neq j$, $i \neq a_i$, $(i \neq a_j$, $i \in [1, 4]$, $j \in [1, 4]$, $i \neq j$, $(x(t_i, t_{ai}) < |R|\text{-}3) \wedge (x(t_{ai,}\ t,) < |R|\text{-}3))$.

B.2 The Neighborhood

The Iterated Local Search phase used the neighborhood structure *SwapRounds.* Let consider *s* be a configuration of the search space. The neighborhood $N : s \rightarrow 2^s$ of *s* is an application such that for all $s \in S, s' \in N(s)$ if and only if *s* and *s'* are different by two rounds. A neighbors *S*
can be obtained by a simple exchange between two rounds.

B.3 The cost function

The cost function is separated in two terms. The first term permits to penalize configurations no satisfying the AtMost constraint. The second term is to penalize those no satisfying the NoRepeat constraint.

First, we defined **occ_hw*(s, t_i, t_j)*** the function that verifies, in a current configuration *s*, if two teams t_i and t_j played in both Home and away in a Round successively.

$(R_{i,j} = R_{j,i} + 1) \vee (R_{j,} = R_{i,j} + 1)$

$$\textbf{occ\_hw}(s, t_i, t_j) = \begin{cases} 1 & \text{if } (R_{i,j} = R_{j,i} + 1) \vee (R_{j,i} = R_{i,j} + 1) \\ 0 & \text{otherwise} \end{cases}$$

The penalty $f_{h,w}(s)$ will be the total number of times that the game is played Home and away successively (in the configuration *s*).

$$f_{h,w}(s) = \sum_{i=1}^{|T|} \sum_{j=i+1}^{|T|} \textbf{occ\_hw}(s, t_i, t_j)$$

Now let consider ***occ-am(s, i,*** R_i***)*** be the occurrence number of team t_i in four rounds from R_i (R_i, R_i+1, R_i+2, R_i+3) successively either home or away.

$$\textbf{con}(s, i, R_i) = \begin{cases} 1 & \text{if } \textbf{occ-am}(s, i, R_i) > 3, R < |R|-3 \\ 0 & \text{otherwise} \end{cases}$$

The penalty $f_{con}(s)$ will be the total number of times the teams play more than three times in four rounds successively.

$$f_{con}(s) = \sum_{i=1}^{|T|} \sum_{Ri=0}^{|R|-3} \textbf{con}(s, i, R_i)$$

The cost function is then defined by the sum of the weights of the two penalties constraints NoRepeat (noted NP) and AtMost (noted AM):

Cost (S)= weight (NP)* $f_{h,w}(s)$ ***+weight (AM)**** $f_{con}(s)$

Solving the CSP problem is searching for a configuration of zero cost. We Noted that we do not deal with the penalty constraint DRRT because we started with initial solution that satisfies the DRRT constraints and we choose a neighborhood structure (SwapRounds) which guarantees that the space research maintain this constraint.

B.4 The ILS algorithm for TTP:

The ILS for TTP is sketched in Algorithm 1.

Algorithm 1: The ILS method.

Require: a SC configurations (satisfied DRRT constraint), max-mov is the maximum number of iterations

Ensure: a feasible configuration CSC(satisfied AtMost, NoRpeat and DRRT constraints)
1: s0← a SC configuration
2: s∗←local search(s0)*;*
3**: for** I ← 0 to max-mov **do**
4: **if** cost (s) ≠ 0 **then**
5 s′← perturbation(s∗)
6: s∗′← local search (s′)
8: **if** (cost (s∗′) < cost (s∗)) **then**
s∗← s∗′
10: **end if**
12: **end if**
13:**end for**
14:**Retutn** the CSC configuration.

C. Phase 3: Stochastic Local Search for TTP

The stochastic local search is a local search method that combines diversification and intensification strategies to locate a good solution. The intensification phase consists in selecting a best neighbor solution. The diversification phase consists in selecting a random neighbor solution. The diversification phase is applied with a fixed probability *wp>0* and the intensification phase with a probability *1-wp*. The process is repeated until a certain number of iterations called *maxiter* is reached.

C.1 The Neighborhood structures :

We make use of three neighborhood relations:

N1: SwapRounds ***(s, R_i, R_j)*** is a simple move that swaps two rounds ***(R_i, R_j)***. There is $O(n^2)$ possible moves. **N2: SwapHomes*(s, t_i, t_j)*** is a move that swaps the home/away roles of teams t_i, and t_j. That means if team t_i plays home against team t_j at round R_k, and away against t_j at round R_l, *SwapHomes* ***(s, t_i, t_j)*** is the same schedule as *s*, except that now team $\boldsymbol{t_i}$ plays away against team $\boldsymbol{t_j}$ at round R_k, and home against t_j at round R_l**N3: SwapTeams** ***(s, t_i, t_j***) is a move that swap the plan of two teams $\boldsymbol{t_i}$ and $\boldsymbol{t_j}$ (except the round where $\boldsymbol{t_i}$ plays against $\boldsymbol{t_j}$.

C.2The SLS algorithm for TTP

The SLS for TTP is sketched in Algorithm 2.

Algorithm 2: The SLS method.

Require: a TTP instance, *maxiter*, *wp*
Ensure: an optimal schedule S for TTP

1: Create an initial configuration (**CS**) verifying the DRRT constraint;
2: Apply Iterated Local Search on CS to obtain a CSC configuration verifying the three constraints
S← solution returned by the Iterated Local Search method
3: Create the totality of the neighborhoods of (S) by applying the *inspiration* technique (N1 structure);

// Generate an initial configuration for SLS with RK
4: create a list (*list*) of movements (R_i, R_j) where the cost(*S*) is equals to zero
5: Generate a random initial solution (CSC) according the code RK , (R_i, R_j)← RK (list); s← Swap Rounds(s, R_i, R_j).

6: **for** I← 0 to maxiter **do**
7: r ⟵ random number between 0 and 1;
8: **if** (r < *wp*) **then**
9: Create the $Movr\ [R_i, R_j].\ by$ neighborhood structure N1
10: random selection of a move (R_i, R_j) (where the cost (*s*) is equals to zero.)
11: s← SwapRounds (s, R_i, R_j)
12: **else**
13: Create the *NBgames list* by neighborhood structure N2
14: Select the best move (t_i, t_j); s← SwapHomes(s, t_i, t_j)
15 : Create the $Movr\ [t_i, t_j].\ by$ neighborhood structure N3
14: Select the best move (t_i, t_j); s← SwapTeams (s, t_i, t_j) ;
17: end if.
18: end for.
19: Return the best solution found

---

3 Experiment

This section gives some experimental results. The source code is written in Java and runs on a Core 2 Duo (1.60 GHz) with 2 GB of RAM.

A. Benchmarks

The proposed method is evaluated on various benchmark problems commonly used in experimental tests taken from the web site NL instance, circular distance instance, Super Instance and Galaxi Instance. The data set used includes 17 instances which are: CON4, CON6, CON8, CON10, CON12, CON14, CON16, NL4, NL10, NL12, NL14, CIRC4, CIRC6, CIRC8, CIRC10, Galaxy 4 and SUPER4. Table 1 gives the numerical results found by our approach. We give the CPU time in second, the best and the average solution found by our method. We give the best know solution for each instance.

B. Parameter Tuning

The adjustment of parameters of the proposed algorithms is fixed by an experimental study. The fixed values are those for which a good compromise between the quality of the solution obtained by the algorithm and the running time of the algorithm is found. The SLS parameters are: the maximum number of iterations= 10000, wp =0.3. The ILS the maximum number of iterations= 20000.

C. Numerical results

In this section, we give the numerical results found by the proposed approach. First, we give in Table 2 the results find by the ILS. The first column gives the number of teams |T|, the second column the number of necessary moves to obtain a

feasible configuration verifying the three constraints The third column gives the CPU time in second to obtain the feasible configuration.

Table 2. Results found by ILS

$\|T\|$	#Mov	Time (s)
6	5	0,096
8	9	0,2
10	11	8,00
12	92	38,53
14	87	78,210
16	201	201,66
18	285	252,35
20	3401	301,29
22	3828	283,60
24	3859	1008,11
26	4115	2563,20
28	7112	1899,22
30	7522	2144,74
32	8001	4055,96
34	89775	5062,45

Table 3. Results found by SLS

Instances	Time(s)	Best known	Result found SLS		Gap%
			Best	Average	
CON4	0,10	**17**	**17**	17	0%
CON6	19,80	**43**	**43**	43	0%
CON8	30,22	**80**	**80**	80	0%
CON10	89,29	**124**	**124**	125	0%
CON12	104;75	**182**	**182**	184	0%
CON14	214,44	**252**	**252**	254	0%
CON16	411,99	**327**	336	338	2,67%
NL4	53,23	8276	8276	8276	0%
NL6	145,21	23916	23916	24122	0%
NL8	524,15	39947	40621	42234	1,65%
NL10	984,73	59583	61193	62711	2,63%
NL12	1635,60	111248	120655	127856	7,79%
NL14	2452,93	188728	206274	231785	8,50%
NL16	5316,34	261687	308413	322394	15,15%
CIRC4	94,65	20	20	20	0%
CIRC6	201,32	64	64	64	0%
CIRC8	431,96	130	140	144	7,69%
CIRC10	866,11	242	272	287	11,02%
Galaxy 4	1245;60	416	416	416	0%
Galaxy 6	108,14	1365	1365	1394	0%
SUPER4	897,89	63405	63405	63405	0%

4 Conclusion

In this paper, we proposed a three-phase search method for solving the traveling tournaments problem (TTP) in sport scheduling. The method is implemented and

evaluated on several benchmark problems with various sizes, and compared with the best known solutions and other techniques for TTP. The experimental results are very promising. The proposed approach provides competitive results and finds solutions of a higher quality. We aim in future to implementing an evolutionary approach for TTP that handles only feasible configurations.

References

1. Allison C. B. Guedes Celso C. Ribeiro, (2010) A heuristic for minimizing weighted carry-over effects in round robin tournaments, Journal of Scheduling 2010.
2. Anagnostopoulos. A, Michel. L, Van Hentenryck.P, and Vergados. Y., (2003), A Simulated Annealing Approach to the Traveling Tournament Brown University, Box 1910, Providence, RI 02912 A Preliminary version of this paper was presented at the CP'AI'OR'03 Workshop. Current Address: University Of Connecticut, Stoors, CT 06269.
3. Bean. .J.C, (1994), Genetics and random keys for sequencing and optimization, In ORSA journal of computing, Vol 6, pp 154-160.
4. Easton. K, Nemhauser. G, and. Trick. M. A, (2001), The travelling tournament problem: Description and benchmarks. In T. Walsh, editor, Principles and Practice of Constraint Programming, volume 2239 of Lecture Notes in Computer Science pages 580–585. Springer, Berlin, 2001.
5. Fabrécio N. Costa S an Urrutia C Ribeiro C., (2012), An ILS heuristic for the traveling tournament problem with predefined venues, in Annals OR, 194 (1), pp: 137-150 (2012).
6. Fleurent. C and Ferland.J.A (1996), Genetic and hybrid algorithms for graph coloring, G. Laporte, I. H. Osman et P. L. Hammer (Eds.), Special Issue of Annals of Operations Research, Metaheuristics in Combinatorial Optimization, vol 63, pp: 437-464, 1996.
7. Irnich S, (2009). A new branch-and-price algorithm for the traveling tournament problem. Technical Report, OR-01-2009, RWTH Aachen; 2009.
8. Luca di Gaspero and Andrea Schaerf, (2006), A Composite-Neighborhood Tabu Search Approach to the Traveling Tournament Problem ; Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica Universit a di Udine via delle Scienze 208, I-33100, Udine, Italy ; May 8, 2006.
9. Nitin S. Choubey A, (2010), Novel Encoding Scheme for Traveling Tournament Problem using Genetic Algorithm ; Department of Computer Engineering,M.P.S.T.M.E., NMIMS-deemed-to-be-University, Shirpur Campus, Shirpur, Dhule, Maharashtra, India, IJCA Special Issue on "Evolutionary Computation for Optimization Techniques" ECOT, 2010.
10. Hof P, Post G, Briskorn D, (2010), Constructing fair round robin tournaments with a minimum number of breaks. Operations Research Letters , vol 38 (2010) , pp: 592–596
11. Ryuhei Miyashiro , Tomomi Matsui, (2006) Semi-definite programming based approaches to the break minimization problem, Computers & Operations Research, vol: 33 (2006) pp: 1975–1982.
12. Rasmussen RV, Trick MA, (2006), the timetable constrained distance minimization problem. In: Beck J, Smith B. editors. Integration of AI and OR techniques in constraint programming for combinatorial optimization problems. Lecture notes in computer science, vol. 3990. Berlin: Springer; 2006. pp: 167–81.
13. Werra. D. (1988), Some models of graphs for scheduling sports competitions. Discrete Applied Mathematics, 21:47–65, 1988.

Parallel Sampling-PSO-Multi-Core-K-means Using Mapreduce

Abdelhak BOUSBACI[1] and Nadjet KAMEL[1,2]

[1]LRIA, Computer Science Department, USTHB Algiers, Algeria
[2]Computer Science Department, Faculty of Sciences, UFAS Setif, Algeria

abousbaci@usthb.dz, akamel@usthb.dz

Abstract. Clustering is partitioning data into groups, such that data in the same group are similar. K-means is the most used clustering algorithm because of its implementation simplicity and efficiency. Many clustering algorithms are based on the K-means algorithm aiming to improve execution time or clustering quality or both of them. Improving clustering quality can be done by an optimal selection of the initial centroids using for example meta-heuristics. Improving execution time can be performed using parallelism. In this paper, we propose a parallel hybrid K-means based on Google's MapReduce framework for the parallelism and the PSO meta-heuristic for the choice of the initial centroids. The results proved that using a network of machines to process data improves the execution time and the clustering quality.

Keywords:K-means; PSO; Sampling; Shared memory; MapReduce.

1 Introduction

Clustering is partitioning data into clusters such that the similarity between objects of the same cluster is high and the similarity between objects of different clusters is low.

The main problem of clustering is obtaining the optimal configuration of clusters and keeping a good execution time.

In the literature, many clustering algorithms have been proposed using different techniques [1], [2], [3], [4].

The k-means algorithm [5] is the most used for its simplicity and efficiency. Its complexity is $O(n * K * I * d)$ where n is the number of data points, K the number of clusters,I the number of iterations and d is the data dimension.

To improve the clustering quality of K-means algorithm various works have been proposed to hybridize it with meta-heuristics, genetic algorithm, particles swarm algorithm... etc. such as in [6], [7], [8], or using sampling process [9] or both such as in [10].

To improve the execution time of K-means, many solutions were proposed, from the optimization of the algorithm itself to the use of the parallelism. The parallelism can be done using two major methods. The first method is the use of a network with many connected machines (master-slave architecture) [11], [12], where clustering

algorithms are processed on a cluster of computers. The second method is the use of shared memory parallelism [13], [14].

To improve clustering quality and execution time, we propose in this paper, to parallelize the Sampling-PSO-K-means algorithm presented in [10].

We propose to parallelize this algorithm using a network of machines managed by MapReduce framework [15]. Using many machines means using many CPUs, and, as we know, nowadays all computers are equipped at least with Dual-core CPUs. If we use a simple clustering algorithm, only one physical core per machine will be used. Therefore, the solution is to avail all cores of a CPU resorting to parallelize the process on each machine.

This paper is presented as follows: Section 2 introduces all the works related to our proposition. Section 3 explains our proposed method. Finally, section 4 is about the implementation and evaluation of our algorithm. In the last section we present our conclusion and perspectives.

2 Related work

2.1 Parallel clustering

Parallelism using shared memory

Many works have been proposed in this class using different approaches [13], [14], [16], [17]. In [13], the authors proposed a solution based on messaging communication between the processes. Other works in this field proposed to use multi-agent systems instead of simple threads like in [17].

In [16] the authors propose an approach based on the use of multi-cores processors and their cores to parallelize clustering algorithms. They parallelized K-means and K-nodes algorithms to cluster gene expressions. The authors chose shared memory parallelism to avoid network communication and to be able to simulate a master-slave architecture. However, the inconvenient of this kind of parallelism is the concurrence in data access. To avoid this problem, using data locks is necessary, but at the same time it can create deadlock problems. The authors proposed McK-means, which is a parallelization of K-means based on shared memory. The proposed solution avoids deadlocks problems over of achieving an improved execution time.

In McK-means algorithm, K-means is parallelized by calculating simultaneously the minimum distance and the centroids update.

Thus, in this approach the initial data set is split into subsets, and nearest centroid search is performed in an individual thread for each subset. This is the parallelization of minimum distance search. On the other hand new centroids calculation is done by a thread for each one. This is the parallelization of the centroids update.

For the minimum distance search, the number of thread is equal to the number of available physical cores, but for new centroids calculation, the number of threads is equal to K, where K is the number of clusters.

Parallelism using machines' network

This is based on using many machines to implement the parallelism when the processes use a large amount of data. Several methods were proposed for this, such in [18] and [19]. The common factor between these works is that authors proposed their own approach of how the machines communicate and how the parallelism is defined. This requires advanced network programming skills. To avoid this, many authors, such as [11] and [12], opted for the MapReduce framework [15].

In its simplest form, MapReduce is a two-step process: a Map step and a Reduce step. In the Map step a master node divides a problem into a number of independent parts that are assigned to map tasks. Each map task processes its part of the problem and outputs results as key-value pairs. The Reduce step receives the outputs of the maps, where a particular reducer will receive only map outputs with a particular key and process them. Since map tasks are independent they can be run in parallel similarly to reduce tasks, which can be completed after the map tasks are completed.

In [20], authors used many connected machines to improve the execution of the K-mean algorithm. They proposed ParC Algorithm which can be summarized in three main steps:

1. Partitioning the initial Data on machines set.
2. Apply a clustering algorithm on each machine using its data partition to find local clusters(*β-clusters*).
3. Merge the clusters found in the previous step that overlap together to get the final clusters configuration.

2.2 A Sampling-PSO-K-means algorithm

The Sampling PSO-K-means algorithm [10] is based on hybrid PSO-K-means algorithm [7]. Authors propose to improve it by sampling the initial data set before being processed by PSO algorithm. It is realized by dividing data into S subsets and applying K-means on each subset. After that, each subset is represented by only its centroids. This step will reduce the global amount of data and keep only the most representative data samples. This process ensures a better execution time and improves the efficiency of the PSO algorithm, meaning an upswing of the entire approach.

This algorithm can be summarized into the following four steps:

1. Select S sub-samples in the initial data.
2. Apply K-means on each sub-sample until convergence.
3. Use the resulted centroids from precedent step with PSO algorithm considering them as swarm particles (*S* particles).
4. Apply K-means on the whole data using the initial configuration obtained from the PSO algorithm.

In this work, authors proposed to use sampling to reduce the amount of data because of the sensitivity of PSO algorithm to large data sets.

We aim to propose an approach to parallelize the Sampling-PSO-K- means algorithm. Our proposal is based on the approaches presented in [16] and [20].

3 Proposed approach: Parallel Sampling-PSO-Multi-Core-K-means Using MapReduce

In this section, we present our contribution: parallel sampling-PSO-McK-means Using Mapreduce.

Our initial objective was to use many machines parallelize the Sampling-PSO-K-means algorithm. On the other hand, as mentioned previously, we aim to exploit CPUs cores. Local parallelism is applied to the Sampling-PSO-K-means algorithm and transform it into Sampling-PSO-Multi-Core-K-means algorithm. Before giving the details of the approach, we give an overview on how it works.

We can describe this process into the three following steps:

1. Distribute data on the cluster nodes.
2. Apply sampling-PSO-McK-means.
3. Merge results from the network machines.

The larger the data set is, the more the PSO algorithm needs iterations to get an optimal solution. With the sampling method in [10] the size of the data set is reduced and PSO can reach an optimal solution with a smaller number of iterations, which means a shorter execution time. Therefore, we proposed to reduce the initial data set in our approach by dividing it on several machines before starting to apply the Sampling-PSO-K-means algorithm. After that, the results are merged.

Sequential processes are executed on many machines in parallel. Since the used CPUs have many cores, using a sequential program on them does not exploit all their potential. So, we propose to use local parallelism instead of the global network parallelism. In Sample-PSO-k-means, K-means algorithm is used twice. At the first time, it is used in the sample step, and then in the final clustering step. Thus the local parallelism will be focused of K-means. This parallelism is done through the following steps:

1. Data instances are separated on P partitions according to the number of available physical cores. and a thread is created for each part (P threads).
2. Each of the P threads is charged to calculate the minimal distance of its partition.
3. Other K threads are created and assigned to the K cluster; K is equal to the number of centroids.
4. Each of the K threads is responsible of calculating the centroid of its own cluster after each update.

This approach avoids the deadlock problem by using a software transactional memory.

The process of our approach can be defined as follows:

1. Initial data is divided on the set of machines.
2. On each machine the following steps are applied using its own data partition:
    a. Data is divided again on S sub-partitions.
    b. A sampling algorithm is applied on each data partition using McK-means.

c. Results from sampling algorithm are used as initial data of the PSO algorithm.
d. Apply McK-means algorithm using the results obtained from PSO algorithm as initial centroids.

3. Results from all the network are sent to a single machine, where cluster of each configuration that overlap in data space are merged together to get the final configuration.

4 Implementation

We implemented our approach, and we tested it on two synthetics numerical multidimensional data sets [21], where data values are in the interval [0.0 , 1.0].

4.1 Data codification

There are three main data components: data set instances, centroids, clusters for k-means algorithms, the particles and the swarm for PSO algorithm.

Each **data instance** is represented by a multidimensional vector: $I_i=(a_{i1},a_{i2},a_{i3},\ldots,a_{im})$ where i the i^{th} data instance.

Each **cluster** is a set of data instances with a variant size. It is represented by a vector of many instances:

$C_i = (I_{i1},I_{i2},I_{i3},\ldots)$ with $i \in \{1,2,3,\ldots,K\}$ where K is the number of clusters.

A **particle** is represented by a vector of K centroids where K is the chosen number of clusters and a centroid is a data instance from the data set. It is represented by a vector of K centroids:

$P_i=(C_{i1},C_{i2},C_{i3},\ldots,C_{ik})$ with $i \in \{1,2,3,\ldots,n\}$, where n is the size of the swarm. A swarm is represented by a vector of n particles.

For the sampling step, the data set is split into many sub-sets, and each one is represented by a vector as follows:

$E_i = (I_{i1},I_{i2},I_{i3},\ldots I_{iM})$ where M is the sub-set size and $i \in \{1,2,3,\ldots,n\}$.

5 Evaluation

Evaluating clustering results means determining how compact the clusters are. To do this, we use the same formula(1) to calculate a particle fitness in PSO algorithm [7].

This formula allows calculating the average distance between the instances of a cluster and its centroid. The smaller this value is the better is the clustering quality. This formula is defined as follows:

$$F = \frac{\sum_{i=1}^{k} \left\{ \frac{\sum_{j=1}^{n_i} d(o_i, i_{ij})}{n_i} \right\}}{K}$$

Where i_{ij} represents the j^{th} data instance of the i^{th} cluster; o_i is the centroid of the i^{th} cluster; $d(o_i,i_{ij})$ is the distance between the data instance i_{ij} and the centroid o_i; n_i is the number of data instances in the cluster C_i and K is the number of clusters.

Distance between two instances is calculated using Euclidean distance.

6 Experimentation

To evaluate the improvement of the hybrid algorithms,we implemented the K-means algorithms. We implemented the Sampling-PSO-K-means algorithm to compare it with our proposed algorithm, the Sampling-PSO-Multi-core-K-means algorithm to prove the efficiency of using threads and the gain in the execution time, the Sampling-PSO-K-means using MapReduce algorithm to demonstrate the improvement brought by using machine's network without multithreading and finally our approach the Sampling-PSO-Multi-core-K-means using MapReduce.

To evaluate our approach we used a cluster of 3 machines. Each node is equipped with a Dual Core CPU and a 2 GB RAM. The cluster run on Linux Ubuntu 10.04. For the framework MapReduce implementation, we used Hadoop 1.2.1 an open source distribution of MapReduce.

The experiments were done on two synthetics data sets [21], DataSet 1 which contains 1400 rows of 8 dimensions and DataSet 2 which contains 10125 rows of 15 dimensions.

The two data sets contain synthetic numerical data, all the values are in the range [0.0 , 1.0].

6.1 Parameters

Many parameters must be defined for each algorithm. The parameters of the Sampling-PSO algorithm are: Inertia factor, confident coefficient at its best position, confident coefficient at its neighboring, number of iterations of PSO, number of particles and number of stamps. We fixed them as mentioned in [10]. The other parameters are the ones of Multi-Core K-means. We cite the number of clusters, the number of iterations and the number of used threads. The latter was fixed in our case to two because we use a Dual-Core CPU. The best number of thread is equal to the number of physical cores [16]. For the number of iterations and number of clusters, we used and tested different values according to the used data sets. For the two data sets we tried many values of the number of clusters using K-means algorithm until getting the best values: $k=3$ for the first data set and $k=5$ for the second one. The number of iterations was defined for each algorithm apart by doing many tests. Finally we get the maximal number of iterations which is *25* and we apply it on all the algorithms for the final tests.

The parameters cited above are summarized in Table 1.

Table 1. Algorithm parameters

Parameters	Value
Number of clusters	3-5
Number of iterations in K-means	5-25
Number of threads	2 (dynamic according to the used CPU)

6.2 Results

To analyse the obtained results we chose the clustering quality and the execution time.

Clustering quality

To evaluate the clustering quality we use the formula 1 to calculate the fitness value of each algorithm. Table 2 shows the results obtained on two different data sets using the parameters presented in Table 1.

Table 2. Algorithms performances

	K-means	Sample-PSO-K-means / Sample-PSO-MC-K-means	Parallel Sample-PSO-MC-K-means using mapreduce
DataSet1	0.3537	0.2193	0.2101
DataSet2	0.4592	0.2691	0.2451

As we can see in the Table 2, the test of the clustering quality was done only for one of the Sample-PSO-K-means algorithm and Sample-PSO-MultiCore-K-means, because both of them give the same results and the only difference between them is the use of threads to improve execution time.

The results show that the Parallel Sample-PSO-MC-K-means using MapReduce algorithm give the better fitness value. Both of Parallel Sample-PSO-MC-K-means using MapReduce and Sample-PSO-MC-K-means algorithms use sampling and PSO algorithm to get a better starting centroids configuration, but the fitness value is better when we use Parallel Sample-PSO-MC-K-means with MapReduce with a similar number of iterations for the both algorithms. This is due to the data partitioning. By dividing the data on the network nodes, we decrease its size on each one. Therefore, PSO algorithm can reach an optimal solution after a given number of iterations. On the other hand, in [22] authors showed that a large data set needs more iterations to converge than a small one does. Our results confirm that. We can see that the improvement of the fitness value is more important on the DataSet2 than on DataSet1, because DataSet2 is larger than DataSets1 so it take more benefits from our approach than DataSet1 does.

In the following, we discuss the convergence of the Parallel Sampling-PSO-MC-K-means using MapReduce and Sampling-PSO-MC-K-means algorithms. In this test, we focus on the observation obtained from the DataSet2. Results are presented in figure 1.

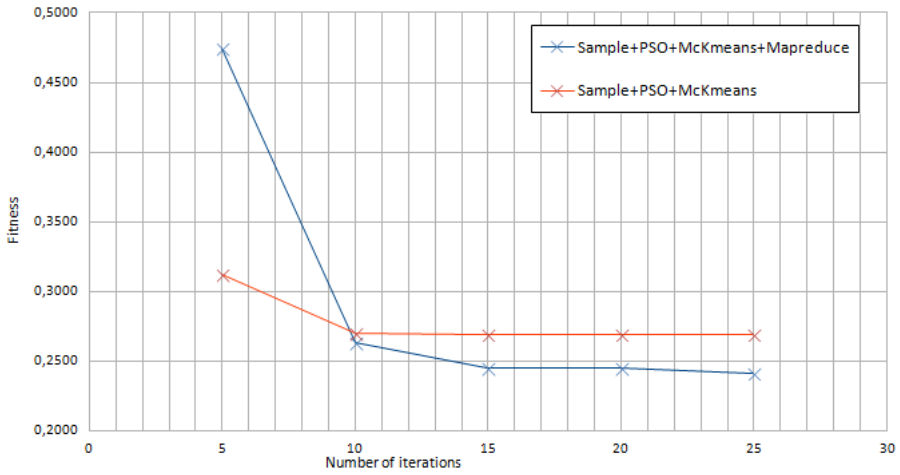


Fig. 1. Convergence graph

We notice that the Sampling-PSO-MC-K-means algorithm get a better fitness after 5 iterations.

Before reaching the 10^{th} iteration, the Parallel Sampling-PSO-MC-K-means algorithm using MapReduce gives a better solution than the Sampling-PSO-MC-K-means algorithm.

After the 10^{th} iteration, the Sampling-PSO-MC-K-means algorithm keeps almost the same fitness value and converges definitively in the 15^{th} iteration. On the other side, Parallel Sampling-PSO-MC-K-means algorithm using MapReduce keeps converging after the 10^{th} iteration until the 25^{th} one.

At the end, we can highlight that the Parallel Sampling-PSO-MC-K-means algorithm using MapReduce algorithm converges to a better solution than the Sampling-PSO-MC-K-means algorithm does.

Execution time

We aim to demonstrate the efficiency of the proposed solution on the execution time parameter. We implemented four algorithms: Sampling PSO K-means, Sampling PSO McK-means, Sampling PSO K-means using MapReduce and the Sampling PSO McK-means using MapReduce algorithm. In this experimentation we used the same parameters in Table 1.

The results are presented in Table 3.

Table 3. Execution time for the implemented algorithms

Algorithm	Execution time (seconds)
Sampling PSO K-means	70,87
Sampling PSO McK-means	52,94
Sampling PSO K-means using MapReduce	24,16
Sampling PSO McK-means using MapReduce	21,98

The results show that the Sampling PSO K-means algorithm has the longest execution time. By applying multi-threading to it (Sampling-PSO-McK-means), we get an important improvement in the execution time.

We observe that with Sampling PSO K-means algorithm using MapReduce, the execution time is faster relatively to the first algorithm. This is do to the fact that the dataset size is reduced by distributing it on the cluster's machines. Thus, every machine will process a smaller amount of data in a smaller time.

Finally, the Sampling PSO McK-means algorithm using MapReduce gives the best execution time by benefiting from the use of many machines and fully exploiting their CPUs.

7 Conclusion

In this paper, we proposed an approach to improve the algorithm presented in [10]. We proposed to distribute the data on many machines and process it in parallel. Afterward, to exploit network nodes, we used multi-threading on each machine to entirely exploit their CPUs.

This approach improved the results of the work presented in [10]. Partitioning data on many machines reduced its amount, and allowed to PSO algorithm to get a solution in a lower number of iterations. At the same time K-means needed less iterations to converge to its best solution due to the data amount reduction.

Instead of clustering quality improvement, we instead considerably improve the response time.

8 References

1. Goil, S., Nagesh, H., Choudhary, A.: Mafia: Efficient and scalable subspace clustering for very large data sets. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (1999) 443-452.
2. Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis. Volume 344. John Wiley & Sons (2009).

3. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. Volume 96. (1996) 226-231.
4. Guha, S., Rastogi, R., Shim, K.: Cure: an efficient clustering algorithm for large databases. In: ACM SIGMOD Record. Volume 27., ACM (1998) 73-84.
5. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Volume 1., California, USA (1967) 281-297.
6. Kwedlo, W., Iwanowicz, P.: Using genetic algorithm for selection of initial cluster centers for the k-means method. In: Artificial Intelligence and Soft Computing Springer (2010) 165-172.
7. Cui, X., Potok, T.E.: Document clustering analysis based on hybrid pso+ k-means algorithm. Journal of Computer Sciences (special issue) 27 (2005) 33.
8. Ahmadyfard, A., Modares, H.: Combining pso and k-means to enhance data clustering. In: Telecommunications, 2008. IST 2008. International Symposium on, IEEE (2008) 688-691.
9. Bradley, P.S., Fayyad, U.M.: Refining initial points for k-means clustering. In: ICML. Volume 98., Citeseer (1998) 91-99.
10. Kamel, N., Ouchen, I., Baali, K.: A sampling-pso-k-means algorithm for document clustering. In: Genetic and Evolutionary Computing. Springer (2014) 45-54.
11. Ene, A., Im, S., Moseley, B.: Fast clustering using mapreduce. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2011) 681-689.
12. Sun, Z.: A parallel clustering method study based on mapreduce. In: 1st International Workshop on Cloud Computing and Information Security, Atlantis Press (2013)
13. Kerdprasop, K., Kerdprasop, N.: A lightweight method to parallel k-means clustering. International Journal of Mathematics and Computers in Simulation 4(4) (2010) 144-153.
14. Stoffel, K., Belkoniene, A.: Parallel k/h-means clustering for large data sets. In: Euro-Par99 Parallel Processing. Springer (1999) 1451-1454.
15. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Communications of the ACM 51(1) (2008) 107-113.
16. Kraus, J.M., Kestler, H.A.: A highly efficient multi-core algorithm for clustering extremely large datasets. BMC bioinformatics 11(1) (2010) 169.
17. Chaimontree, S., Atkinson, K., Coenen, F.: A multi-agent based approach to clustering: Harnessing the power of agents. In: Agents and Data Mining Interaction. Springer (2012) 16-29.
18. Guerrieri, A., Montresor, A.: Ds-means: distributed data stream clustering. In:Euro-Par 2012 Parallel Processing. Springer (2012) 260-271.
19. Hammouda, K.M., Kamel, M.S.: Models of distributed data clustering in peer-to-peer environments. Knowledge and information systems 38(2) (2014) 303-329.
20. Ferreira Cordeiro, R.L., Traina Junior, C., Machado Traina, A.J., Lopez, J., Kang, U., Faloutsos, C.: Clustering very large multi-dimensional datasets with mapreduce. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2011) 690-698.
21. School of Computing, U.o.E.F.: Clustering datasets (2012).
22. Davidson, I., Satyanarayana, A.: Speeding up k-means clustering by bootstrap averaging. In: IEEE Data Mining Workshop on Clustering Large Data Sets. (2003).

Reformulation's Optimization of Unleaded Premium Gasoline with Ethanol

[(*)]Karima Ibelaidene[1,2], Djamel El Hadi[2]

[1] Sonatrach. Activité Amont Division Technologies et Développement,
Avenue du 1er novembre Boumerdes. Algeria
[2] Laboratoire d'analyse fonctionnelle des procédés chimiques, Département de Chimie
Industrielle, Université de Blida1, Route de Soumâa
B.P. 270 09000 Blida1- Algérie

[(*)] k.ibelaid@yahoo.fr

Abstract. To improve the quality of Algerian gasoline by suppressing lead, this study is conducted for reformulating, via ethanol blend, an unleaded premium gasoline to reach international specifications, and thus to preserve environment. For this purpose, gasoline with blending stocks used in refinery has been prepared. Furthermore, the optimization of different blending stocks' contributions in the mix leads us to maximize the octane number.
On the observation of the reformate quality (low octane number), the final reformulated gasoline did not reach the octane number as dictated by international specifications (RON > 95). These new findings based principally on octane number property, give us the possibility to evaluate the performance of catalytic reforming units of the refinery

Key words: Gasoline – reformulation - environment – ethanol – optimization- Research Octane Number.

NOTATIONS

CO	carbon monoxide
d	density
Er	relative error in percentage
RON	Research Octane Number
NO_x	nitrogen oxide
NO_2	nitrogen dioxide
R^2	coefficient of determination
RVP	Reid Vapor Pressure

1 Introduction

The most required fraction issued from crude oil is gasoline, mainly with the advance of auto industry which requires more and more quality allowing, thus, to obtain the best performances of the vehicle.

The improvement of vehicle efficiency is due to the increasing of the octane number of our fuels. It is the most important characteristic which measures fuels quality. To increase the octane number of fuel, leaded additives are used. This type of fuel is called "leaded high grade petrol".

Vehicles provided with leaded gasoline release pollutants. The main ones are: the carbon monoxide (CO) which, exposed to oxygen, are quickly transformed into nitrogen dioxide (NO_2). NO_2 combined with water's atmosphere forms acid rains. Under ecological pressure, fuel industries must modify fuel formulations to produce cleaner high grade petrol. Producing unleaded gasoline becomes a necessity which leads refiners to make big efforts to keep the octane number at a satisfactory level. A biased solution has been found: adding oxygenated products to reduce the concentrations of polluting components and to increase the octane number. This study aims to determine a mathematical model in which octane number varies with different blend stocks composition of gasoline. This model enables us to: 1) find the factor which influences strongly the octane number and 2) determine the blending stocks proportions to be mixed in order to maximize the octane number. An experimental validation of the model has been carried out.

2 Experimental Results

2.1 Characteristics of Blending Stocks

First, gasoline marketed by Algiers refinery used as a reference, has been characterized. The blend stocks used in the reformulation are:

- SR, Straight Run gasoline, obtained from atmospheric distillation of crude oil;
- Reformate, obtained from catalytic reforming;
- Light solvent, obtained from atmospheric distillation of crude oil.

The main characteristics of premium gasoline marketed by Algiers refinery as well as used blending stocks are shown in table 1.

Table 1: Characteristics of premium gasoline (reference) and blending stocks

Characteristics	Reformate	SR Gasoline	Light solvent	premium gasoline
Density at 15°C	0,7574	0,6387	0,7180	0,7548
RVP (kPa)	35,9	144,7	25,9	66,3
RON	90	60	72	96
Refractive index at 20°C	1,4340	1,3645	1,4026	1,4365
Aniline point (°C)	< 26	53	58	19
Sulfur content (ppm)	< 20	< 20	< 20	0,1
Boiling range (°C)	41,7-185,5	20,3 - 95,7	66,2 - 129,5	40 - 190,4
Residue in (vol. %)	1,1	0,5	1	1
Losses in (vol. %)	1,4	4,1	33,2	1,5
n-paraffins (wt %)	16,535	53,8	38,2	19,1
Iso paraffins (wt %)	44,575	40,4	25,7	41,2
Naphthenes (wt %)	7,904	4,9	2,9	12,2
Aromatics (wt %)	29,866	0,9	0	27,9
Lead content (g/l)	0	0	0	0,4

2.2 Gasoline reformulation:

At a start, we have prepared mixtures by using various feed stocks proportions after calculation of the most important characteristics such as: density (d), Reid Vapor Pressure (RVP), Research Octane Number (RON) considering these characteristics as additive properties.

Gasoline preparation is the most important step in this work: five mixtures have been carried out by the use of various bases gotten from Algiers refinery. To these bases, we added ethanol, the concentrations of which are used in the different reformulations. The results are shown in table 2.

According to the obtained results, it has been noticed that not only does ethanol presents a better value of octane number, but also a higher vapour pressure.

Once we added ethanol to reformulated gasoline, we retained five mixtures of composition. They are shown in table 3. The main characteristics of reformulated gasoline are shown in table 4.

Table 2 Characteristics of ethanol

Characteristics	Values
Density at 15°C	0,7931
RVP (kPa)	155
Refractive index at 20°C	1,3678
RON	106
Purity (wt %)	98,8

Table 3: Composition of the Reformulated Gasoline (Vol%).

Blend stocks	Gasoline 1	Gasoline 2	Gasoline 3	Gasoline 4	Gasoline 5
Reformate	80	73	70	71	75
SR Gasoline	13	20	17	23	10
Light Solvent	2	2	8	1	10
Ethanol	5	5	5	5	5

Table 4: Physicochemical Characteristics of the Reformulated Gasoline

Characteristics	Gasoline 1	Gasoline 2	Gasoline 3	Gasoline 4	Gasoline 5
Density at 15°C	0,7434	0,7350	0,7367	0,7368	0,7433
RVP (kpa)	56,7	67,2	61,4	61,7	53,9
RON	87	85,6	84,2	84	84
Refractive index at 20°C	1,4479	1,4260	1,4281	1,4228	1,4488
Aniline point (°C)	< 26	< 26	< 26	< 28	< 28
Sulfur content (ppm)	< 20	< 20	< 20	< 20	< 20
Boiling range (°C)	35,5 - 183,2	36,9 - 180,6	34,5 - 181,4	37 - 181,1	38,3 - 181,3
Residue in (vol. %)	1	1,1	1,1	1	1,1
Losses in (vol. %)	0.7	1,2	1,5	1,2	1,9
Hydrocarbon types:					
n-paraffin hydrocarbons (wt. %)	20,830	18,454	21,617	22,893	19,967
Isoparaffin hydrocarbons (wt. %)	37,978	53,050	40,109	41,037	40,384
Naphthenic hydrocarbons (wt. %)	4,608	10,582	10,820	10,114	11,128
Aromatic hydrocarbons (wt. %)	28,355	15,681	25,425	23,497	26,432

3 Calculation Part

3.1 Suggested Correlation:

In order to estimate the octane number of reformulated gasoline, a mathematical model is suggested. It is based on the use of multiple linear regressions. Octane number is an affine function of the composition X_J (volumetric fraction) of the different blending stocks. It leads to the following model:

$$RON = 94{,}85X_1 + 70{,}64\ X_2 + 57{,}60\ X_3 + 11{,}74\ X_4\ . \qquad \textbf{(1)}$$

J = 1 (Reformate)
J = 2 (SR Gasoline)
J = 3 (light Solvent)
J = 4 (Ethanol)

To determine the various constants accuracy with this correlation, we associate the value of the relative error by cuts with determination coefficient R^2; both are defined by the following formula:

$$Er\ (\%) = 100(\ \sum \left| RON_{exp} - RON_{th} \right| / RON_{exp})/n\ . \qquad \textbf{(2)}$$

$$R^2 = \sum (RON_{th} - \xi)^2 / \sum (RON_{exp} - \xi)^2\ . \qquad \textbf{(3)}$$

Where,
$RON_{exp,}$ is the experimental or the measured value of RON, $RON_{th,}$ is the estimated value of RON,
n, is experimental number of points
ξ, is given by the formula 4:

$$\xi = (\ \sum RON_{exp}\ ) / n\ . \qquad \textbf{(4)}$$

Values of calculated statistical parameters (Er (%) = 0.604% and R^2 = (0.998) show that the suggested model is characterized by a very high accuracy. This allows the adjustment of the octane number value according to the mixture compositions.

3.2 Optimization of the Reformulation:

According to all encountered difficulties in practice, the linear programming provides a vast framework to treat a large variety of the linear optimization problems, where linear programs are undoubtedly the most frequent.

Refinery produces premium fuels. They are composed of a mixture of several blending stocks with different quality. Their characteristics are previously mentioned in table 1.
Produced fuel must have a high octane number. Therefore, our goal is to maximize this characteristic. The gasoline properties must respects standards as defined by industrials. The mains are:

- RVP should not exceed, in summer, 60kPa;
- density at 15°C varies from 0,735 to 0,785;
- Aromatics contained in gasoline mixture should not exceed 35%;
- And RON ≥ 95

The sum of all contents must be equal to 1 ($X_1 + X_2 + X_3 + X_4 = 1$)
with $X_1 \geq 0$, $X_2 \geq 0$, $X_3 \geq 0$ and $X_4 \geq 0$.

We seek to maximize the function RON = f (X_1, X_2, X_3, X_4) by solving the following linear system:

$$\begin{cases} 35{,}95X_1 + 144{,}7X_2 + 25{,}9X_3 + 155X_4 \le 60 \\ 0{,}7574X_1 + 0{,}6387X_2 + 0{,}7180X_3 + 0{,}7931X_4 \ge 0{,}735 \\ 0{,}7574X_1 + 0{,}6387X_2 + 0{,}7180X_3 + 0{,}7931X_4 \le 0{,}785 \\ 29{,}866X_1 + 0{,}9X_2 + 2{,}9X_3 \le 35 \\ X_1 + X_2 + X_3 + X_4 = 1 \\ X_4 = 0{,}05 \\ X_1 \ge 0;\ X_2 \ge 0;\ X_3 \ge 0 \end{cases}$$

The resolution of the optimization program shows that if we mix the blending stocks with the proportions: 78,3% of reformate; 16,6% of SR gasoline and 5% ethanol, we get a maximum value of octane number equal to 86,65. To check these results, we proceed by experiments. We mix the selected blending stocks by using the preceding proportions. The measured octane number of this gasoline has the value of RONexp = 86.

3.3 Evaluation of the Results

In this work, it has been found a correlation according to which the octane number varies in function of used blending stocks' proportions. The mathematical method used has enabled us to determine gasoline composition corresponding to the maximum value of octane number of gasoline. The obtained results are not satisfactory since we could not reach the required octane number (min 95) as defined by standards.

In view of the experimental checking, it has been observed that the experimental value of the octane number (RON_{exp} = 86) is closed to the one found through optimization (RON_{th} =86,65), with a relative error of 0,76%.

4 Conclusion

This work is regarded as an ambitious project which aims at the reformulation of premium unleaded gasoline with respect of the international specifications so as to take part in the international market as well as the protection of the environment.
Since lead is classified as a major pollutant, its toxicity is much more important than that of benzene and aromatics. In our study, we have used bases of Algiers refinery by introducing ethanol as a new base.

To conclude, we can summarize this work as follows:

Reformulated gasoline presents an octane number relatively far from the required norms. This is mainly caused by bad quality of refinery reformate with an octane number value of 90. This problem is justified by the ageing and poisoning of catalyst

reforming unit. We have also tried to find a correlation with a mathematical method setting the variation of octane number in terms of gasoline compositions blend.

The optimization results have successfully been tested and can be improved by using other bases available in our refineries.

In future research, we suggest two solutions to improve octane number value which is caused by reformate quality: either, to rehabilitate the reforming unit or to dope the pool gasoline with weak concentration of heavy aromatics (10 to 12%).

References

1. Guibet, J.C., Martin, B. and Montagne, X.: Gasoline composition effects on exhaust emissions. 13th World Petroleum Congres, Buenos Aires. (1991)
2. Guibet, J.C., Faure, E. : Carburants et moteurs. Technologies. Energie. Environnement, (Eds.) Technip (1997)
3. Guibet, J. C.: Utilisation des produits organiques oxygénés comme carburants et combustibles dans les moteurs, Première partie: Aspects techniques de l'utilisation sur moteurs, Revue de l'Institut Français du Pétrole, Vol.36, N°5, (1981)
4. Wauquier, J.P. : Pétrole brut – Produits pétroliers – Schémas de raffinage. (Eds.) Technip (1994)
5. Nocca, J.L., Forestière, A. and Cosyns, J.: IFP's new technologies for formulated gasolines. NPRA Meetting, San Antonio (1990)
6. Nocca, J. L., Forestière, A. et Cosyns, J. : Nouvelles technologies IFP pour la reformulation des essences. Revue de l'Institut Français du Pétrole, Vol.49, N°5. (1994)
7. Douaud, A., : Carburants et moteurs de demain : Le moteur à essence : quel challenge pour les années 2000. Congrès international SIA Paris (1994)
8. Ibelaid, K.. : Reformulation de super carburants sans plomb par ajout de composes oxygénés. Thèse de Magister en Génie Chimique, E.N.P. Alger, Algérie (2005)
9. Mathieu, D., Phan-Tan-Luu R. : Planification d'expériences en formulation : criblage. Techniques de l'Ingénieur, traité Génie des procédés, J2240.
10. Mathieu, D., Phan-Tan-Luu, R. : Planification d'expériences en formulation: optimisation. Techniques de l'Ingénieur, traité Génie des procédés, J2241.

Hybrid ACO and Tabu Search for Web Information Foraging

Yassine Drias and Samir Kechid

LRIA, USTHB,
USTHB, BP 32 El Alia, Bab Ezzouar Algiers, Algeria

ydrias@usthb.dz, skechid@usthb.dz

Abstract. In this paper, we propose two ACO algorithms for information foraging (IF) on large-scale data sets. The first novelty in this work is the design of a model to adapt ACO approaches and even other meta-heuristics to IF. The second one resides in the hybridization of ACO approaches with taboo search in order to achieve more efficiency. The designed algorithms were implemented for comparison purposes. Experiments were conducted on MedlinePlus, a benchmark dedicated for research in the domain of Health. The results are promising either for those related to some Web regularities and for the response time, which is very short and hence complies with the real time constraint.

Keywords: Information Foraging, Web intelligence, ACO, Tabu Search, MedlinePlus.

1 Introduction

Nowadays, Web intelligence tends to evolve in permanence. A major concern is the development of the Information Foraging (IF) paradigm. It consists in surfing the Web to get useful information under a time constraint. This issue may appear at first glance simple and with no major interest. However, its importance is stimulating nowadays the Web users as the latter is in an incessant growing and the human ability to explore the astronomical amount of data on the web is relatively very limited. Besides, tackling such issue is very welcomed in domains like business, finance, health and science. The potential users not only will spend less time getting the localization of the needed information but they can even get it in real time.

1.1 Information foraging background

Web mining is the analysis and discovery of data, documents and multimedia information existing on the Web. It includes studies of features such as the structure of hyperlinks, Web usage statistics and search of the Web sites contents. Studies on the

structure including hyperlinks allow the detection of the pages that have authority on a certain topic. Web Usage statistics used techniques to discover patterns in the log files of users. Meanwhile searching the Web contents aims at getting the closest possible information needs for users by presenting the most appropriate Web pages.

The concept of information foraging (IF) shares the same goal as Information Retrieval (IR), which is information search. While IR uses a complex process (indexing and matching), IF consists in navigating from one page to another for the same purpose.

The present study deals with simulating IF by taking into consideration the complex structure of the Web. We designed a system for IF and a tool for helping users to undertake IF with efficiency.

1.2 Ant Colony Optimization

As we know, real ants apply a stigmergetic way of communication by the use of a hormonal secretion called pheromone. In fact, the ants deposit on ground pheromone when moving and tend to choose the path which has the greatest amount of pheromone. To search food, ants take different directions, but those choosing the shortest path will reach the food more quickly. When they return, the pheromone on the shortest path will be stronger and will attract the successive ants to take this path [1].

In ACO algorithms, artificial ants imitate this way of communication by using artificial pheromone, which is some numerical information saved on the states of the search space of the problem to solve. The first ACO algorithm known in the literature is the ant system AS. It has been proposed by Dorigo [1], and has various extended versions like the Max-Min AS called MMAS, the rank based version [3] (ASrank) and the Ant Colony System ACS[4]. The approach has gained a large reputation since it has solved with success many combinatorial optimization problems.

The recent work that addresses IF with meta-heuristics is BSO-IF, with bee swarm approach [2]. We think that ant system is more appropriate for foraging, this is why we are interested by ACO for Web navigation.

Motivated by the success and the power of this meta-heuristic and knowing that very few if none of heuristic search techniques have been devoted to investigate information foraging problem, we designed two ACO algorithms, namely AS-IF and ACS-IF for exploring this domain.

The algorithms we designed were tested on MedlinePlus, an online benchmark about health.

2 Related works

Liu in his talk at IJCAI'03 [5], suggests new directions for research in the new field of Web Intelligence (WI) that has emerged a decade ago from artificial intelligence and information technology. The main goal of WI is to develop theories and technologies towards using optimally the connectivity of the Web.

In [6] and [7], the authors proposed an agent-based model for IF and validated it using empirical Web log datasets. They consider Web topology, information distribution and interest profile in building a Wisdom IF agent. They found out that the unique distribution of agent interest leads for regularities in Web surfing and that Web regularities are interrelated. They also undertook an interesting study on three categories of users according to their interest and familiarity with the Web: A random user with no intention in surfing, a rational one with an objective but with no familiarity on surfing and a recurrent user who is familiar with the Web and who has a goal for surfing. The result is that independently from the kinds of users, the regularities of user surfing are the same, which means that the user ability for predicting the surfing chain is predominant.

Strong regularities in Web surfing were also studied by Huberman et al. [8] from the theoretical point of view. They proposed a model for studying surfing behaviours and the experiments they held showed common surfing behaviours. The study conducted by Huberman et al. in [9] shows that the Web pages are distributed over the sites according to a universal power law, which is an example among the other strong regularities.

In [10], Ibekwe-SanJuan describes in a clear and nice manner the recent concepts, methods and applications of text mining. The chapter on Web mining contains a rich documentation on notions that concern the new developments of the Web technologies such as Web topology, sites popularity, sites ranking and propagation of metadata to co-links.

Chi and Pirolli in [11] introduce the concept of social information foraging and its understanding. They explored models for social IF and focused on the importance of the benefits of cooperative foraging.

Bees Swarm Optimization for Web Information Foraging was developed in [2]. The authors simulate the human behaviour while searching for information using artificial bees.

From the above literature review, we remark that the rare existing papers related to IF offer new ideas on how to develop theories and technologies about IF, which means that the area is still in an early stage. The theory developed by Pirolli et al. is an important advancement and can stimulate future works in the field. On the other hand, the agent-based model proposed by Liu et al. can be considered as the commencement of future IF technologies.

The original contribution of the present paper consists in developing an approach based on Ant Colony Optimization (ACO) for IF.

3 Authorities Mining using Aunt Colony Optimization

A colony of artificial ants is launched to seek authorities according to their behaviour that guarantees finding the richest places of the target. The authorities mining takes into account the Web topology and the user interest profile. The group of ants works for one user, which is different from simulating a group of usersbehaviour. The latter

issue was studied in [11]. Whereas in our case, the cooperative feature is handled in the implementation by the ants.

The search starts from the website's homepage, then guided by their senses, the ants are directed to the goal and after several generations, they find the right outcomes.

4 Algorithm AS-IF

In this section, we present the ant system algorithm called AS-IF designed for information foraging. Let us first start with the description of the problem modelling. Based on the natural antsbehaviour for finding food from a very large geographical space, the Ant System (AS) algorithm simulates this process for finding optimal solutions from a huge set of potential solutions. The ants move from the hive to a food source and when reaching the latter, they alert their congeners by means of a stigmergetic communication asking them for help to transport the food to the hive. According to animal psychology, this communication is performed thanks to the pheromone that the ants deposit on ground to orient the congeners to the place containing an important amount of food.

4.1 Solutions encoding

The search space for the colony of ants will be the MedlinePlus database. In AS, ants encapsulate solutions and a solution for our application is a surfing path. So ants will seek Web surfing paths that end with authorities. The adaptation of AS to IF is called AS-IF and is outlined in Algorithm 1.

4.2 Pheromone table and probabilistic decision rules

The ant algorithm includes several ant generations, each generation is composed of *NbAnts* ants. Two structures are needed to compute the ant algorithm, a table named *Phero* to store the pheromone amount yielded by the ants each time it builds a solution and a table called *sol* to save the best solution found by each ant. *phero[k]* corresponds to the pheromone amount associated to the document found by ant *k* and *sol[k]* is the best solution determined by ant k. Ants will construct new solutions using these structures, which represent a means of communication between the artificial ants. The tables are updated at each generation of ants. Besides, two variables namely *best* and *bestsol* are used to save respectively the best solution found during one generation and the best solution computed since the beginning of the process. Each ant starts building a solution from an initial solution *s* generated randomly. It then constructs a solution using a stochastic process. The ant chooses a solution from its neighborhood, with a probability computed as follows:

$$P(k) = \frac{phero[k]}{\sum_{j=1}^{j=NbAnts} phero[j]} \quad (1)$$

Algorithm 1 AS-IF

```
Input: N: MedlinePlus database; user interest;
Output: bestsol: A surfing path ending with an authority;
 1: procedure AS-IR
 2:     for k=1 to NbAnts do phero[k]= 0.1;                ▷ pheromone initialization
 3:     end for
 4:     select at random a solution s from N;               ▷ a surfing path namely s
 5:     best := bestsol := s;
 6:     for i=1 to MaxIter do
 7:         for  k=1 to NbAnts do
 8:             generate a random initial solution s;
 9:             sol[k] = s;
10:             s' := build_AS (s) ;
11:             sol[k] := s;                          ▷ update the best solution of k
12:             update the online pheromone phero[k] using formulas (2) and (3);
13:             if f(s') > f(best) then then best := s';    ▷ f is the similarity function
14:             end if
15:         end for
16:         if f(best) > f(bestsol) then bestsol := best;
17:         end if
18:         apply online-update of pheromone;
19:     end for
20:     return (bestsol);
21: end procedure
```

The neighbourhood N_i of a document is the set of documents attached with $term_i$ in the inverted file as shown in Figure 4. All these documents are neighbours because they share at least one term between them. One neighbourhood contains a huge number of documents where the AS algorithm is launched.

The pheromone information is initialized with a small value equal to 0.1 in order to simulate the fact that initially the real ants deposit a very small amount of pheromone on the ground when starting their space exploration. During the search, the pheromone amount, which represents the importance of the surfing path, will be computed and associated to each surfing path found by the ants. The AS-IF framework considers one sub-collection of the website pages corresponding to one user interest. It is outlined in procedure AS-IF().

4.3 Updating the pheromone

The strategies of updating pheromone simulate the evaporation of natural pheromone followed by a production of this chemical substance. The evaporation phenomenon gives rise to rule (2) where the empirical parameter ρ belongs to the interval [0, 1] and simulates the evaporation rate. For online update performed at each generation of ants, the pheromone added is calculated according to rule (3) whereas for the offline update rule (4) is applied. Recall that *bestsol* is the best solution found during the previous iterations and *best* is the best solution of the current iteration.

$$phero[k] = (1 - \rho) * phero[k] \quad (2)$$

$$phero[k] = phero[k] + \rho * f(s) \quad (3)$$

$$phero[k] = phero[k] + \rho * f(bestsol)/f(best) \quad (4)$$

4.4 Building and improving a solution

Each ant performs the task of exploring the best surfing path in a sub-collection. The method designed for this aim is described through the procedure called build_AS. It merely chooses a solution from its neighbourhood with probability *p* defined in formula (1).

Algorithm 2 Build_AS

Input: N_s: surfing path s;
Output: best_s : a surfing path with a better authority;

```
1: procedure BUILD_AS(var s)
2:     draw at random a page i from N_s;
3:     compute p = P(k) using formula (1);
4:     generate at random a number r from [0,1];
5:     if r > p then s = i;
6:     end if
7:     best_s = tabu-search(s);
8:     return(best_s)
9: end procedure
```

After its construction, each solution undergoes an improvement of its quality by applying the procedure tabu-search (Algorithm 3). In the tabu search procedure, the considered neighborhood is the one previously described. The intensification phase starts when the number of iterations without improving the solution quality reaches some limit. After applying the intensification strategy, a diversification technique is launched by choosing the less recently used moves and thus directing the search to new regions of the space. We use the variables best_s, which is the best solution found by the tabu process. And in order to set the stop condition we need also the variable, namely no-improve, which informs about the number of successive iterations without improvement of *best_s*. The variable *max-no-improve* is the maximum number of iterations without improvement before starting the intensification process. The procedure *tabu-search* outlined below calls both the procedure *neighbor(s)* that returns the best nearest neighbor of *s* which is not tabu nor satisfies the aspiration criterion and the procedure *update-t-length()*, which updates the tabu list length.

Algorithm 3 Tabu-search

```
Input: N_s: the surfing path;
Output: Best_s: the best surfing path;
 1: procedure TABU-SEARCH(var s)
 2:     best_s = s;
 3:     no-improve = 0;
 4:     while (not stop condition) do
 5:         s := neighbor(s);
 6:         update-t-length();
 7:         if (f(s) ≤ f(best_s)) then no-improve =: no-improve + 1;
 8:         end if
 9:         if (no-improve = max-no-improve) then
10:             no-improve := 0;
11:             best_s := Intensification (s);
12:             if f(s) > f(best_s) then best_s := s;
13:             end if
14:             best_s := diversification (best_s);
15:         end if
16:     end while
17:     return (best_s);
18: end procedure
```

5 Algorithm ACS-IF

The second designed ACO algorithm is the ant colony system (ACS) version. Its main difference with the previous one resides in the design of the probabilistic decision rules and the procedure of building solutions which is called build_ACS (Algorithm 4).

$q_0$is a tunable parameter and the pseudo-random-proportional rules are computed using respectively the probability of (5) or (6).

$$P(k) = \begin{cases} 1 \text{ if } f(k) = argmax_{j \in V_{sol[k]}} (phero[k]^{\alpha} heur[j]^{\beta}) \ for k = 1 \ldots NbAnts \\ 0 \text{ else} \end{cases} \quad (5)$$

$$P(k) = \frac{phero[k]^{\alpha} heur[j]^{\beta}}{\Sigma_{j=1}^{j=NbAnts} phero[j]^{\alpha} heur[j]^{\beta}} \quad (6)$$

The probability P(k) in (6) is computed using the quantity of pheromone and a heuristic function. $\alpha$and$\beta$are empirical parameters and control respectively the importance of these two components. The heuristic part is calculated as follows:

$$heur[k] = max_{j \in V_{sol[k]}} f(j) \quad (7)$$

In other words, the ant decides stochastically to consider the best solution found in the neighborhoods of the solutions being treated during the current iteration when $q \leq q_0$and a document drawn at random otherwise, unless the computed probability of formula (6) is greater than a generated random number *r*.

Algorithm 4 Build_ACS

Input: N_s: surfing path s;
Output: best_s : a surfing path with a better authority;

```
1: procedure BUILD_ACS(var s: solution)
2:     generate a random variable q;
3:     if (q ≤ q0) then
4:         let s be the surfing path with the best authority;
5:         put s in the taboo list;
6:     else
7:         choose a non taboo surfing path s randomly;
8:         compute probability P(k) by rule (6);
9:         generate a random value r ∈ [0, 1];
10:        if P(k) > r then s := argmax_{j∈V_k} f(j);
11:        end if
12:        put s in the taboo list;
13:    end if
14:    best_s := tabu-search(s);
15:    return (best_s)
16: end procedure
```

6 Experimental Results

6.1 Description of the real-world Benchmark

Extensive experiments were performed on an on-line medical database provided by the U.S. National library of Medicine called *MedlinePlus*. It provides information on over 900 diseases, health conditions and wellness issues. Our experiments deal only with health topics described in an XML file that includes pages describing medical topics. The data is available at http://www.nlm.nih.gov/medlineplus/xml.html. We worked on the version of the 1st January 2015 where the number of nodes was equal to 1905 with a total volume of 27 MB. Each topic is specified by a title and contains the following elements: an URL, an identifier, the language of the topic (English or Spanish), the date of its creation, eventually tags specifying among others, topic synonyms, translation to other languages, a full summary, related topics, which are internal links to similar topics and external sites.

Only links to related topics are exploited because they belong to the database. External sites are ignored as they direct to pages outside the database.

6.2 Experimentations and Results

Our work was implemented in Java Eclipse help System Base, version 2.0.2 on a PC with an Intel core I5-3317U Processor (1.70 GH) with 4 GB RAM. Figure 1, Figure 2 and Table 1 exhibit the results of the experiments of selecting the the optimal parameters, which are shown in Table 2

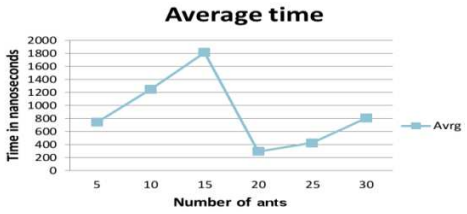


Fig.1.*Number of ants vs. time*

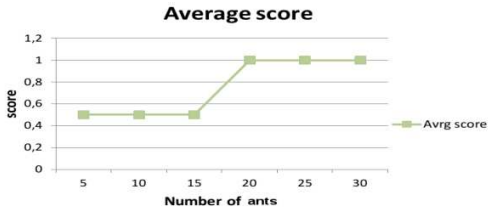


Fig. 2. Number of ants vs. score

Table 1. Choosing α and β parameters

α	**β**	**score**
1	1	1.41
1	2	0.5
2	1	4.94
2	2	$3.43*10^{19}$

Table 2. The parameters that yields the best results

Number of ants	30
Maximum number of generations	35
TL	30
α	2
β	1

The results we focused on are: the authority page (the surfing page with the highest score), its URL, its score, the surfing depth and the surfing time in nanoseconds. They are shown in Table 3.

7 Conclusion

In this work, we proposed a hybrid ACO and tabu search approach for Web information foraging. Unlike previous works, our approach is inspired from nature and biological psychology and adopts an analogy with animals groups hunting, which simulates real-world surfing. The second originality of the present study is in the use of a real-word site server MedlinePlus for the experiments instead of an artificial one.

We implemented the system using Ant Colony Optimization (ACO) on MedlinePlus. This idea is inspired from nature where animals hunt together in a group and rarely alone. The results are promising.

The perspectives for this study are numerous. For the short term, we are thinking about integrating the user preferences in the user interest profile in the surfing model.

Table 3. *Experimental Results for different user interest profiles*

User interest	Relevant Page		Score	Surfing depth	Surfing time
	Title	URL			
Pain, Abdominal	Abdominal Pain	*/abdominalpain.html	1.0	2	122
H5N1	Bird Flu	*/birdflu.html	1.0	2	140
Heart, Diseases	Heart Diseases	*/heartdiseases.html	1.0	5	155
Hypersensitivity	Allergy	*/allergy.html	1.0	1	98
Cancer	Cancer	*/cancer.html	1.0	2	138
Poor, Blood, Iron	Anemia	*/anemia.html	1.0	2	152
High, Blood Pressure, Medicines	High Blood Pressure	*/highbloodpressure.html	0.75	3	1291
Skin, Allergies	Skin Conditions	*/skinconditions.html	0.5	2	844
MCI	Mild Cognitive Impairment	*/mildcognitiveimpairment.html	1.0	4	313
Anorexia	Body Weight	*/bodyweight.html	0.15	20	5641
Recovery, surgery	After Surgery	*/aftersurgery.html	1.0	2	199
Pimples	Acne	*/acne.html	1.0	1	301

* : http://www.nlm.nih.gov/medlineplus

References

1. M. Dorigo, L.M. Gambardella : Ant algorithms for discrete optimization. Artificial Life. 5-3, 137-172, 1999.
2. Y. Drias, S. Kechid : Bees Swarm Optimization for Web Information Foraging. MIKE'14, LNAI 8891, Springer 189-198, 2014
3. B. Bulheimer,R.F. Hartl, C. Strauss : A new rank based version of the ant system, a computational study. Technical report POM -03/97, institute of management science, university of Vienna, 1997.
4. O. Cordon, I. Deviana, F. Herrera, L. Moreno : A new ACO model integrating evolutionary computation concepts: the best-worst ant system. From Ant Colonies to Artificial Ants, ANTS 2000. 22-29, 2000.
5. J. Liu, Web Intelligence: What Makes Wisdom Web? Invited Talk, IJCAI 2003.
6. J. Liu, N. Zhong, Y. Y. Yao, and Z. W. Ras. The Wisdom Web: New challenges for Web Intelligence (WI). Journal of Intelligent Information Systems, Kluwer Academic Publishers, 20(1):5-9, 2003.
7. J. Liu and S. W. Zhang. Characterizing Web usage regularities with information foraging agents. IEEE Transactions on Knowledge and Data Engineering, vol. 16, N°. 5, May 2004, 566-584.
8. B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose. Strong regularities in World Wide Web surfing. Science, 280:96-97, 1997.
9. B. A. Huberman and L.A. Adamic. Growth dynamics of the World-Wide Web. Nature, 410:131, 1999.
10. FideliaIberkwe-SanJuan: Fouille de textes méthodes, outils et applications, Hermès, Lavoisier 2007.
11. E. H. Chi, P. Pirolli: Social Information Foraging and Collaborative Search.HCIC Workshop, Fraser CO, 2006.

Improving Source Selection Process using social profile

Zakaria Saoud and Samir Kechid

LRIA, USTHB,
USTHB, BP 32 El Alia, Bab Ezzouar Algiers, Algeria

zakaria.saoud@live.fr, skechid@usthb.dz

Abstract. In this paper we present a new personalized approach that integrates a social profile in distributed search system. The previous distributed information retrieval research based on the textual information defined new measures to improve the different process in the distributed information system, but neglected the use of the social information. From this point, we propose an approach which exploits the different social entities to: (i) make a query expansion, and (ii) personalize and improve the source selection process in distributed information retrieval.

Keywords : Social information retrieval , Source selection, User profile

1 Introduction

With the increase of the web size, the amount of information covered by a centralized search engine decreases [6].Thus, the centralized Information retrieval is no longer sufficient to satisfy the needs of users. To solve the problems of centralized Information Retrieval, distributed Information Retrieval appeared, which consist of using meta-search engines to increase the research coverage and by combining the results from several centralized search engines. Unlike to the centralized information retrieval, the big potential, in which the distributed information retrieval can achieve it, is the possibility of getting the information from different sources [21].The distributed information retrieval gave birth to two major problems: (1) the source selection and (2) the result merging. In this paper, we are interested in the source selection problems. The large amount of results returned by meta-search engines is a great disadvantage, and from there the custom meta-search engines have emerged. Personalized meta-search engines use profiles of users to filter the results and return the relevant documents that better meets to user's needs. Many kinds of personal data can be used for the construction of the user profile such as user manually selected interests [19], search engine history[20],etc. Internet growth and the advent of web2.0gave birth to different types of social networks on a large scale, which are now recognized as an important means for information dissemination [8]. Many social networks are considered associal tagging systems; these systems allow the users to provide annotations (tags) to resources, to give their opinionsabout resources.Severalsocial bookmark-

ing services, such as Flickr[1] and Delicious[2], are considered an online folksonomy services, and their social tagging data, also known as folksonomies [2]. The set of tags can be used as a source of personal data to build the user profile. In this work we propose new personalization approaches that exploit the social relations among items and tags through the use of user profile defined within a social tagging system to make a query expansion and to improve the source selection process, in order to improve the quality of searching of a meta-search engine. The rest of the paper is organized as follows: Section 2 describes the related work. Section 3 details our personalization approach, and we describe the experimental setup and results in section 4. Finally, we conclude our work and list some future work in Section 5.

2 Related Work

2.1 Social information retrieval

Social tagging systems are web-based systems that allow internet users to add, edit, and share bookmarks of web documents. In social bookmarking services, such as Delicious and Flickr, the users can annotate their bookmarks with arbitrary keywords called tags. The collection of a user's tags constitutes their personomy, and the collection of all users' personomy constitutes the folksonomy. A folksonomy is a tuple $\mathbb{F} := (U, T, D, Y)$ where$U = \{u_1, \ldots, u_M\}$ is the set of users,$T = \{t_1, \ldots, t_L\}$ is the set of tags, and $D = \{d_1, \ldots, d_N\}$ is the set of resources or web documents, and Y is a ternary relation between U and T andD, i. e., $Y \subseteq U \times T \times D$, whose elements are called tag assignments[2][11].In our case, the elements ofDrepresentsthe different web resources and are identified by a URL. Users are identified by a user ID. In social information retrieval, many studies have proposed in the context of search personalization. Most of these studies are based on the folksonomy structure.

Schenkel et al.[9]developed a framework for harnessing such social relations for search and recommendation. They created a scoring model that exploits social relations and semantic/statistical relations among items and tags; this scoring model gives a great importance to users who have a high score of friendship strength with the query initiator. The score offriend ship strength is a linear combination of the spiritual friendship strength, the social strength friendship, and the global friendship strength. Rather than item recommendation, our personalized retrieval models, applicable to meta-search engine, is composed of several web search engines to re-rank the lists of search results according to the user profile.

Bender et al. [1] exploited the different entities of social networks (users, documents, tags) and social relations between these entities, to make a query expansion by adding the similar tags to the query keywords, and to make a social expansion to give an advantage to documents tagged by the user's close friends.

[1]Flickr - Photo sharing, http ://www.flickr.com/
[2]Delicious - Social bookmarking, http ://delicious.com/

Hochul et al.[3] developed an approach which uses the links and similarities between the user profiles in the filtering algorithm results. This approach is called collaborative filtering. The principal advantage of this approach is the enlargement of the coverage of research using similar profiles. For example, in the case where P does not obtain satisfactory results for a query, we can then use the most similar profiles to P to enlarge the search and retrieve more relevant results.

Vallet et al. [11] presented a personalization model that exploits folksonomy structure. For this two measures were developed to calculate the relevance between a user profile and a document to re-rank the list of results returned by a search engine.

2.2 Source selection approach

Source selection is a decisive step in the metasearching process [14]; it aims to reduce the number of selected sources for a given query, in order to not lose the search time when there are many sources of information, by selecting only the relevant sources to the user's query.

Several source selection methods have been developed, and they can be classified into two main categories: manual selection methods, and automatic selection methods. In automatic selection, several approaches have been defined.

Si and Callan [22] proposed an approach called CORI (Collection Retrieval Inference network), which consider a collection as a meta-documents, and the selection is made according the similarity between the user query and the source.

Savoy and Rasolofo[13] used a learning method for the selection of servers, which is based on a decision tree. The learning process matches for each server and each query a set of pairs (attribute, value) and the decision is based on the selection or not of the server.

Si and Callan[12] proposed another source selection approach called UUM (Unified Utility Maximization Framework for Resource Selection), where they are based on the estimation of the size of the source, to estimate the number of relevant documents can be contained in the source. The estimated number of relevant documents used for the selection of sources.

Kechid and Drias[4][5] calculated a score for each source. This score combines three measures: 1- The source similarity according to the user interest, 2- The source similarity according to the user query, and 3- The accuracy degree between the source features and the user preferences. Arguello et al [15] presented a source selection approach that combines multiple sources of evidence to inform the selection decision, and they derive evidence from three different sources: collection documents, the topic of the query, and query click-through data.

Hong et al [16] proposed a novel probabilistic model for resource selection process, through combining the evidence of individual sources and the relationship between the sources, to estimates the probability of relevance of information sources.

3 Social Personalization approach

The previous distributed information retrieval approaches based on the textual information, to define new measures for the personalized search, however they neglect the use of the social information, and they didn't exploit the social information to improve the different process in distributed information retrieval. Hence we decided to define a social profile and exploiting it for personalizing and improving the source selection process in distributed information retrieval.

Similar to the studies of Hochul et al [3] and Bender et al. [1], we exploit the folksonomy structure, and we use both the friendship measure and the similarity between two tags to make a query expansion rather than the use of the friendship and the similarity measure directly in the scoring model.

Similar to the studies of Vallet et al [11], we follow the same personalization model but we exploit other relations between the folksonomy entities (users, documents, tags),such as the relation between user and user and the relation between tag and tagto expand the research coverage and to improve the search quality of a distributed information system, particularly, to improve the source selection process.

3.1 User profile definition

The user profile is defined by the user's set of tags; we suppose that unlike the rarely used tags, the most frequently used tags are more relevant and significant to describe the user's interests, hence, we use just the tags that have frequency greater than or equal to the average of user's tags. We note:

$$\text{profile}(u_m) = \left\{ \left(t_i, tf_{u_m}(t_i) \right) \middle| tf_{u_m}(t_i) \geq avgtags_{u_m} \; and \; i \in [1..L] \right\}$$

where:

L: is the number of tags used by the useru_m.

$tf_{u_m}(t_i)$:is the User-based tag frequency, which mean show many times the eru_m use the tagt_i.

$avgtags_{u_m}$: represents the average of all tags used by the useru_m, we calculate this average using the following proposed formula:

$$avgtags_{u_m} = \frac{\sum_{i=1}^{L} tf_{u_m}(t_i)}{tags\_number_{u_m}}$$

where$tags_number_{u_m}$: represent the number of tags used by the user u_m.

The user profile in our approach is used to:

1) Work a query expansion by adding similar tags to the keywords that appear in thequery.

2) Select the most suitable sources according to the user profile.

3.2 Query expansion process

According to Barry Smyth [10], 66% of our new research is similar to those made by our colleagues. Thus we can say that it is important to take into account the social factor, in information retrieval. Based on this idea, for our query expansion process, we propose the use of the profiles of the user's close friends, to find the list of tags that are similar to the query terms. The user query is reformulated by adding the similar tags of the query terms, according to the query initiator's close friends. The new query is weighted by the following formula inspired by Rocchio[7].

$$q^{new}(u_m) = a.q^{old} + \frac{b}{|\overline{T}|}\sum_{\acute{t} \in \overline{T}} \acute{t}$$

With q^{new} is the expanded query,q^{old} is the old query given by the user, a and b are constants, $a, b \in [0,1]$.For example if we want to expand the query "cooking video" according to user "1423", we add the similar tag of the term "cooking" and the similar tag of the term "video", which are respectively the tag "food" and the tag "youtube", so we obtain the new query "cooking video food youtube".

$\overline{T}$: represents the list of similar tags, which will be added to the old query. This listis established by the following proposed formula:

$$\overline{T} = \bigcup_{t_i \in q^{old}} \acute{t}, \acute{t} = \{\acute{t} \in sim(t_i) | TagSim(\acute{t}, t_i) = \max_{sim(t_i)}\}$$

*$sim(t_i)$: refers to the list of tags that are similar to the query term t_i, which belongs to the list of bookmarks of the user close friends. This list is generated as follows proposed formula:

$$sim(t_i) = \{\acute{t} | TagSim(\acute{t}, t_i) > 0 \text{ and } \acute{t} \in bookmarks(friends_{close}(u_m))\}$$

To calculate the similarity between two tags $\acute{t}, t_i$, we use the Dice coefficient measure defined as follows:

$$TagSim(\acute{t}, t_i) = \frac{2 \times df_{\acute{t},t_i}}{df_{\acute{t}} + df_{t_i}}$$

where:

- $df_{\acute{t},t_i}$: is the number of documents which belongs to the list $friends(u_m)$, and that have been tagged by both tags $\acute{t}$and t_i.
- $df_{\acute{t}}, df_{t_i}$: are the number of documents which belongs to the list $friends(u_m)$, and that have been tagged with $\acute{t}$and t_i, respectively.

In our approach, two tags are similar, if they have a high probability to appear together in the same documents. For example, the similarity degree between the tag "technology" and the tag "science" according to user "2611" is calculated as follows:

$$\text{TagSim}(\text{technology}, \text{science}) = \frac{2 \times 4}{44 + 30} = 0.108$$

To get the close friends list of the query imitator, we use our following formula:

$$\text{friends}_{\text{close}}(u_m) = \bigcup_{y \in \text{friends}(u_m)} y, \text{friendship}(u, y) \geq \text{avg\_friendship}(u_m)$$

where:

- $\text{friends}_{\text{close}}(u_m)$: refers the list of close friends of the query initiator u_m, the close friend of user u_m must have a friendship score greater than or equal to the average score of friendship.

- $\text{avg\_friendship}(u_m)$: the average friendship score of the user u_m, which calculated using the following proposed formula:

$$\text{avg\_friendship}(u_m) = \frac{\sum_{y \in \text{friends}(u_m)} \text{friendship}(u_m, y)}{|\text{friends}(u_m)|}$$

*$\max_{\text{sim}(t_i)}$: represent the tag that has the high similarity in the list $\text{sim}(t_i)$.

3.3 Source selection process

Source selection process aims to select the most relevant sources for a given user's query[18]. In our approach we aim to integrate a social profile in the source selection step, and for that we define a score $\text{ScoreSource}_s(u_m)$for each source *s* associated to the user u_m, and based on this score, we select and sort the relevant sources for each user.

The score $\text{ScoreSource}_s(u_m)$ is calculated by combination of both measures $\text{SimSource}_s^{\text{terms}}(u_m)and\text{SimSource}_s^{\text{tags}}(u_m)$as follows:

$$\text{ScoreSource}_s(u_m, q) = (1 - \alpha).\text{SimSource}_s^{\text{Terms}}(q) + \alpha.\text{SimSource}_s^{\text{Tags}}(u_m, q)$$

where:$\alpha \in [0,1]$

The parameter$\alpha$is used to control the influence of the two measures of ty $\text{SimSource}_s^{\text{terms}}(u_m)and\text{SimSource}_s^{\text{tags}}(u_m)$. For example if$\alpha$ is height, the selected sources will have a heights degrees of similarity with the adapted social profile of the query initiator$u_m$andif α=0these sources will be selected according the degree of similarity with the user's query. The other utility of the parameter $\alpha$is tonormalize the global score $\text{ScoreSource}_s(u_m)$, in order to not exceed the interval [0, 1].

$\mathrm{SimSource}_s^{\mathrm{Terms}}(q)$: represent the degree of similarity between the source s and the user's queryq, according to the set of terms of the source documents. This similarity is calculated as follows:

$$\mathrm{SimSource}_s^{\mathrm{terms}}(q) = \frac{\sum_{i=1}^{T} t_i * q_i}{\left(\sum_{i=1}^{T} {t_i}^2\right)^{1/2}\left(\sum_{i=1}^{T} {q_i}^2\right)^{1/2}}$$

With,

s:isthe source; q is the user's query ;

t_i:is the weight of the term i in the source (the set of the first k documents returned by the source);

q_i: is the weight of the term $i$in the query;

T:is the number of terms used in the source. In our approach, the content of a source is represented by the set of its first k returned documents.

$\mathrm{SimSource}_s^{\mathrm{Tags}}(u_m)$: represent the degree of similarity between the source s and the query initiator u_m, According to the set of tags of the source documents. This similarity is calculated using the measure of Noll and Meinel[17] as follows:

$$\mathrm{SimSource}_s^{\mathrm{tags}}(u_m, q) = \sum_{j=1}^{K} \mathrm{tf}(u_m, d_j)$$

withK is the number of returned documents .

d_j : is the document j of the source s.

$\mathrm{tf}(u_m, d_j)$: is the similarity measure of Noll and Meinel, which is defined as follows:

$$\mathrm{tf}(u_m, d_j) = \sum_{\substack{i=1 \\ i \in d_j}}^{i=L} \mathrm{tf}_{u_m}(t_i)$$

where:

$\mathrm{tf}_{u_m}(t_i)$:is the number of times the user $u_m$hasused the tag t_i.

4 Experiments

The evaluation of personalized information retrieval approaches difficult task, because of the absence of personalized relevance judgments. Therefore we have decided to construct a test collection using a social bookmarking dataset and a set of documents downloaded from several search engines to simulate a real distributed environment and to provide the social information.

4.1 Experimental Setup

For our experiments, we used a dataset from the del.ico.us social bookmarking system; this dataset is released in the framework of the 2nd International Workshop onInformation Heterogeneity and Fusion in Recommender Systems (HetRec 2011)[3]. This dataset contains social networking, bookmarking, and tagging information from sets of 1867 users from Delicious social bookmarking system. It contains 69,226 URLs (resources), 1,867 users, and 53,388 distinct tags.
To evaluate our approach in a distributed environment, we considered the 8 following search engines : GOOGLE, YAHOO, BING, BLEKKO, YANDEX, WOW, ASK,AOL. Each search engine represents a source of information, and for each source we have downloaded the top 30 retrieved documents, and each result list returned by a source is used to create the source description.
To examine the benefit of our approach for individual users, we allowed 12 participants to evaluate the returned documents by the selected sources, and each user ran 6 queries. In total, we have tested 72 different queries. These queries are made using the most popular tags in the dataset, in order to increase the proportion of obtaining tagged documents from the search engines to be able to apply our personalization approach, across the most returned documents. To solve the problem of subjectively assessing (relevance judgment) [1], wefollow the method of Bender et al [1], and that by selecting a fictitious profile for each query initiator. In our test, the fictitious profile is extracted from the set of profiles of the social bookmarking dataset del.ico.us, and this profile must contain the greatest sum of tags frequency of the query.

4.2 Experimental results

4.2.1 Results of personalization approach

In this section, we analyze the performance of our personalization approach whenonly the personalization scores are used to select the different sources of information.
For each query we considered the result in various cases, by varying the parameters α and in the interval [0, 1]. We vary the parameters αto evaluate the influence of the use of the social profile, and to evaluate the importance between the two measures of the source selection score. To evaluate our selection approach without using a merging algorithm, we follow the method of Arguelloet al[15]. This method consider that the source s is relevant with a query q,if the source s contain more thantrelevant documents which are preset in top Tof the full-dataset result.The full-dataset is the set of the different documents returned by the whole sources, which indexed by a function into a centralized indexed list. In our evaluation we select between 1-6 sources, and we combine the returned documents by the$k$ selected sources into a single list, then we select the top 30 relevant documents according to the users' judgments. We allow each participant to judges the relevance of the 1, 2,3,4,5 and the 6 selected sources. A precision value is computed for each retrieval session according to the following formula:

[3]http ://grouplens.org/datasets/hetrec-2011/

$$\text{precision} = \frac{number\ of\ relevent\ documents\ in\ the\ k\ selected\ sources}{number\ of\ returned\ documents\ by\ the\ k\ selected\ sources}$$

The various cases obtained by varying the parameters $\propto$ in the interval [0, 1] are described as follows:

- case $1 : \propto = 0$:
  This case means that the social profile is not used ; the relevance of the result is related just to the user query.
- case $2 : 0 < \propto < 0.5$:
  In this case, the relevance of the result is related to the user profile and the user query, but the user query is more significant than the social profile.
- case $3 : \propto = 0.5$:
  In this case, the relevance of the result is related to the user profile and the user query, in an equitable way.
- case $4 : 1 > \alpha > 0.5$:
  In this case, the relevance of the result is related to the user profile and the user query, but the social profile is more significant than the user query.
- case $5 : \propto = 1$:
  This case means that the user query is not used; the relevance of the result is related just to the social profile.

In each case we computed an average precision for the whole queries .Table 1 shows the average precision (at $P@\{1,2,3,4,5,6\}$) values of the personalization approaches for each case.

Table 1. Average precision values of the personalization approach.

k	***Case* 1**	***Case* 2**	***Case* 3**	***Case* 4**	***Case* 5**
1	0.133	0.230	0.230	0.133	0.166
2	0.116	0.200	0.183	0.133	0.133
3	0.155	0.188	0.177	0.177	0.155
4	0.150	0.166	0.175	0.166	0.158
5	0.140	0.160	0.153	0.146	0.146
6	0.127	0.144	0.138	0.133	0.122
Average	0.136	0.181	0.176	0.148	0.146

From this table we can see that the second and the third case, give better results than the first case, which uses the user query and neglects the social profile, and better results than the last case, which uses just the social profile and neglects the user query. Thus we can see the advantage and the utility, when we integrate the social profile who presents the social information with the user query who presents the text information.

The second case gives better results than the fourth case, w ho gives more importance to the social profile than the user query, and better results than the third case, whi-

chuses the social profile and the user query in an equitable way. Therefore we can deduce that the user queryis more significant than the social profile for the relevance of theirsearch results. As a result, we can say that the combination of the social profile with the user querycan improve the source selection process, and gives the best results of the retrieval process when we give more importance to the user query.

4.2.2 Results of query expansion

In this section, we study the performance of the personalization approaches when we apply our query expansion method for each query. To realize that we preferred the use of the popular tags to made the queries, which contain just a single tag, to avoid the changing of the meaning of the queries. After the collection of the new queries, we apply our personalization approachas the previous section. For each query, we have computed the precision values for the 5 and 10 first selected sources. Table4 shows the averages precisions obtained in each case:

Table 2. Average precision values of the query expansion method.

k	***without query expansion***	***with query expansion***
1	0.174	0.181
2	0.153	0.158
3	0.162	0.175
4	0.165	0.169
5	0.149	0.153
6	0.132	0.148
Average	0.155	0.164

From this table we can remark that the query expansion method can improve the personalized approach results. We mention that the query expansion result can vary, depending on the user profile. For example if we want to expand the query "android" for the user "8691", we obtain the query "android mobile", but if we expand the same query for the user "6585", we obtain the query "android ipod" because of the difference between the users' profiles, which forms a distinction between the users' interests, and can affect the query expansion results.

5 Conclusion and future work

In this paper we have defined a new social user profile based on the folksonomy structure. We have also defined a new approach using the social user profile for personalizing and improving the retrieval process in distributed information retrieval. The results obtained shows that the integration of the social profile, in source selection process, improved the relevance of the distributed information retrieval. In addition, the second evaluation indicates that the use of the query expansion process with the

social profile gave good results for the source selection process. In our future work we plan to adapt the user profile according to the query, to make the search more specific. We would like also to evaluate our approach using a large dataset to obtain reliable results.

6 References

1. Bender, M.; Crecelius, T.; Kacimi, M.; Michel, S.; 0001, T. N.; Parreira, J. X.; Schenkel, R. &Weikum, G. (2008), Exploiting social relations for query expansion and result ranking., in 'ICDE Workshops' , IEEE Computer Society, , pp. 501-506 .
2. Hotho, A.; Jäschke, R.; Schmitz, C. &Stumme, G. (2006), 'Information Retrieval in Folksonomies: Search and Ranking', The Semantic Web: Research and Applications , pp. 411-426.
3. Jeon, H.; Kim, T. & Choi, J. (2010), 'PersonalizedInformationRetrieval by Using Adaptive User Profiling and Collaborative Filtering.', AISS 2 (4) , 134-142 .
4. Kechid, S. and Drias, H. (2009). 'Personalizing the Source Selection and the Result Merging Process.' International Journal on Artificial Intelligence Tools (IJAIT 2009) 18 (2) , 331-354.
5. Kechid, S. and Drias, H. (2010). 'Personalised distributed information retrieval-basedagents'. IJISTA 9(1): 49-74 (2010). International Journal of Intelligent Systems Technologies and Applications. Vol. 9, No 1, 2010.
6. Lawrence, S. & Giles, C. L. (1999), 'Accessibility of Information on the Web.', Nature 400(6740) ,107-109
7. Rocchio, Jr., J. J. (1971), Relevance Feedback in Information Retrieval 'The SMART Information Retrieval System' , Prentice Hall, , pp. 313--323 .
8. Saito K., Kimura M., Ohara K., Motoda H. (2010) ' Selecting Information Diffusion Models over Social Networks for Behavioral Analysis ' , European Conference, ECML PKDD 2010, pp.180-195.
9. Schenkel, R.; Crecelius, T.; Kacimi, M.; 0001, T. N.; Parreira, J. X.; Spaniol, M. &Weikum, G. (2008), 'Social Wisdom for Search and Recommendation.', IEEE Data Eng. Bull. 31 (2) , 40-49 .
10. Smyth, B. (2011), Web Search: Social & Collaborative., in Gabriella Pasi& Patrice Bellot, ed., 'CORIA' , Éditions Universitaires d'Avignon, , pp. 3 .
11. Vallet, D.; Cantador, I. & Jose, J. M. (2010), Personalizing Web Search with Folksonomy-Based User and Document Profiles., in Cathal Gurrin; Yulan He; Gabriella Kazai; Udo-Kruschwitz; Suzanne Little; Thomas Roelleke; Stefan M. Rüger& Keith van Rijsbergen, ed., 'ECIR' , Springer, , pp. 420-431 .
12. L. Si, J. Callan. Unified Utility Maximization Framework for Resource Selection. Proceedings of the international conference on Information and knowledge management CIKM washingt on USA 2004, pp 32-41.
13. Savoy J., Rasolofo Y. Recherche d'informations dans un environnement distribué. In actes "Traitement Automatique de la Langue Naturelle, TALN 2000 Lausanne (Suisse), octobre 2000, pp. 317-326.
14. P. Ipeirotis and L. Gravano. When one sample is not enough: improving text database selection using shrinkage.SIGMOD, pages 767–778, 2004.
15. Jaime Arguello, Jamie Callan, and Fernando Diaz. 2009. Classification-based resource selection. In Proceedings of CIKM. 1277–1286.
16. Dzung Hong, Luo Si, Paul Bracke, MichaelWitt, and Tim Juchcinski. 2010. A joint probabilistic classification model for resource selection. In Proceedings of SIGIR. 98–105.
17. Noll, M. G., Meinel, C.: Web search personalization via social bookmarking and tagging.In: Proc. of ISWC 2007. LNCS, vol. 4825, pp. 367-380. Springer, Heidelberg (2007).

18. Markov, I., Crestani, F.: Theoretical, qualitative, and quantitative analyses ofsmall-document approaches to resource selection. ACM Trans. Inf. Syst. 32(2)(April 2014) 9:1-9:37.
19. Pazzani, M., Muramatsu, J., Billsus, D.: Syskill & webert: Identifying interesting web sites. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1. AAAI'96, AAAI Press (1996) 54-61.
20. Paul, J., Gowan, M.: A Multiple Model Approach to Personalized Information Access. (2003).
21. Steidinger, A.: Comparison of different collection fusion models in distributed information retrieval (2000).
22. Callan J., Lu Z., Croft B. Searching distributed collection with inference networks. In the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, ACM-SIGIR'95, p. 21-28, july 1995.

Contextual source selection for federated search in mobile environment

Hamid Benachour and Samir Kechid

LRIA, USTHB,
USTHB, BP 32 El Alia, Bab Ezzouar Algiers, Algeria

Hamid.benachour@gmail.com,kechidsam@yahoo.fr
{hamid.benachour,skechid}@usthb.dz

Abstract. The emergence of new communication technologies like 3G, 4G and the latest generation of smart-phone devices, allows the user to continuously access the Web, at any time and from any location with different devices. The introduction of the user and his environment in the research process is intended to solve the information overload problem and to increase the accuracy of the retrieval system. In this paper, we present a new approach to select the best sources from different heterogeneous sources by exploiting a multidimensional contextual user profile in a mobile environment, that includes the research situation "time, location ...", the device and the user preferences. We present also our approach to define users and sources profiles, using external ontologies and a learning algorithm supervised by users results satisfaction score.

Keywords: federated search, mobile environment, contextuel source selection

1 Introduction

Today the user is facing an extraordinary flow of information on the web; the system used for guiding the user becomes important to reduce the complexity and diversity of the received information. Indeed the information retrieval systems "IRS" selects from a collection of documents, relevant information to meet the user needs expressed generally by a query. Classical search engine use web pages crawlers and save data in a centralized index but, In fact several studies [1] [2] show that the size of unindexed deep web is estimated to 70 %. That's why federated search emerges.In order to better meet user informational needs, current works defines a contextual profile in addition to the initial request, incorporating supplementary dimensions depending on the nature of the problem, in fact the great evolution of mobile technologies need to take account several new contextual factors.
Most works define a federated search also known as distributed information retrieval systems "SRID" [4], as three sub-processes, the representation and the selection of the sources , and merging results into one list. This paper aims to build a profile which include the user situation and context in a retrieval information process over various

sources. More precisely, we are interested by the sources selection phase, for such needs we have enriched the context of the user by the spatio-temporal and events aspect as well as devices characteristics to define the user situation when he querying the retrieval information system.This new mobile environment has upset the traditional research practices, [4, 5] show that queries are shorter and ambiguous, and users informational needs depends more on context and environment according to [6]. With this complexity, our primary focus in this paper is to include the user and her current situation in the source selection process.
This paper is organized as follows: the second section presents the problems related to the sources selection of federated search in a mobile environment and a state of the art of related work. The third section presents our approach to define the context of a mobile user. The fourth section presents our source profile definition. The fifth section presents our source selection approach. This article ends with a conclusion that summarizes our contributions and presents our research perspectives.

2 State of the art and related work

Distributed retrieval information in mobile environment is across between several research axes

Sources selection. There are a considerable work dedicated to distributed information retrieval system "SRID" in the literature, More precisely, to the three sub-problems, known as sources representation and selection, and merging the results in one list [7]. The main purpose of the sources selection is to reduce the cost of research by reducing the number of servers to query. Most existing selection methods adopt a ranking sources method, according to their degree of relevance. The first approaches like CORI (Collection Retrieval Inference Network) [10],LWP [13] treat each sources as one big document. Some of these algorithms use a probe request to got additional information used to calculate the final score of each server.Another sources selection algorithms was developed, such as [12] based on probabilistic model, or [14] where the authors have exploited the language model to evaluate the divergence between the language model of the query and the sources. Authors in DTF [15], CRCS [16] represent documents individually and use their ranks and relevance to scoring sources, CRCS use a centralized sample index document, and exploit the k best sample documents to calculate final sources scores. While DTF takes into account the quality and cost of servers retrieval information system RSI.Another works category based on a learning algorithms as [17], which use a decision tree to learn a source selection model. in [18] authors use a requests categories obtained by a classification of ODP[1] directory results, while in [19] the authors exploit user feedback to made a classifier that estimates relevant sources.Recent work has suggested other selection method in various research area. In Taily [20] the authors use a cooperative selection method based on a language model, the query's scores represented by Gamma

[1] http://www.dmoz.org/

distribution in each sources and documents with the highly scores are selected. In [21] the authors use a decision tree learned using tourism website to exploiting it as a post-filtering in federated contextual suggestion system. In [3] the authors use a sample index weighted according to [10], a set of document is recovered using the user query, all documents are referenced to their original collection, each collection is shown as a graph where each document is a point with a score and a rank, the final documents scores represents the area between the plot and the x-axis.

Retrieval information in mobile environment. A first category of work focused on the adaptability of research systems on mobile devices considering their constraints. Several approaches propose adaptive methods for a research results visualisation [22], and geographic search interfaces usability [8]. Another category of work addressed the facilitate entering queries [23] and exploiting suggestion, [24] and auto-completion [25] to help the user to express his informational needs. Other work has exploited the user context to adapt and improve the retrieval search system in a mobile environment. In [26] the authors propose a multidimensional profile for a mobile user, which includes the location and time in addition to user's cognitive context, a case-based reasoning "CBR" approach is adopted to select the appropriate profile and reorder search results. In [27] the authors propose a personalised system that alternates between two states moving or stationary according to user's activity, each state has kinds of content information and special interface corresponds to the time, location, and a set of personal information filled before.In [11] the authors use available information on social networks to set preferences and user interest, as in [26] the authors in [9] propose a concepts representation of the interests and preferences of the user based on an ontology, and guided by mining search results and their click through as validation method, SVM is then used to re-ranking future search results.

3 Defining the contextual user's profiles for a federated search in mobile environment

Our user's situation model use a high-level information based on external ontologies to better reflect the users needs in a semantic scale, and to annotate the data crawled form different mobile sensors with a concepts. As is shown in Figure 1. our context model can be represented by a set of situations US, and a range of domains interest UI, and a set of preferences device DP and as wall as a set of user preference UP. We can distinguish global profile that includes "users situation US, device preferences DP" and local profile that represents "user interest UI, user document preferences UP", where each global user profiles contains one or many local profiles.

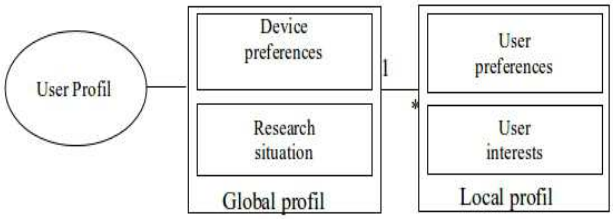


Figure 1: User profile of our approach

3.1 Global user profile

The global profile instance represented by vector of two dimension "users situation, device preferences" noted $< US, DP >$

User research situation. We describe the user situational context by a vector of four dimensions, <Time, Location, Event, Move speed> which we define as $USi =< T_i, L_i, E_i, MS_i >$

*Time.*For a good representation we have reused the ontology OWL Time[2], and extend it with classes that defining more closely the real situation of the user. Day time can be represented by morning, noon, afternoon, evening and night, each one are represented by work or rest time periods. Week time, is represented by work or rest days.

*Location.*We have reused LinkedGeoData [3] that uses a spatial data collection OpenStreetMap[4] to create a large spatial knowledge base. This ontology consists of more than 1 billion nodes and the resulting RDF data includes about 20 billion triples.The data is available according to the Linked Data principles and interlinked with DBpedia[5] and Geo Names[6]. This allowed us to move from the physical location of the user "gps coordinates" to a semantically meaningful location.

Move speed. The informational needs depends on user activity, we exploit the instantaneous speed based on Doppler shift, and we represent it with two states moving or stationary.

*Events.*We have exploited an external ontology linkedevents[7], wish have extended with temporal and spatial classes LinkedGeoData and OWL ontologies-time and the concepts of general ontology DBpedia to express the topic of the event.

[2] http://www.w3.org/TR/owl-time/
[3] http://linkedgeodata.org
[4] http://www.openstreetmap.org/
[5] http://dbpedia.org
[6] http://www.geonames.org/
[7] http://linkedevents.org/ontology/

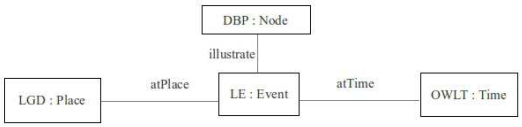


Figure 2: Graph of the event ontology

Device preferences.A mobile user can use several types of device such as a pc at work, a laptop at home, a tablet in the garden. The device preferences can influence the situation and the informational needs, user may not consult an result not suitable to his device interface. We represent the device preferences by the following vectors:

– <Device resolution, Format compatibility, Connection speed>

The vectors instances are automatically updated upon arrival of a new non existing instance. For the same situation we calculate the average of preference devices used in the research process, so for each vector instance DP_i^n we calculate the moving average of L similar past situations US that included the n preferences, by the following formula:

$$P(DP_i^n) = (\frac{1}{L}) \sum_{i=0, i \neq j}^{L} US_i^n \quad (1)$$

Where, DP_i^s is the n instance of device preferences, L the depth of the average to fix in implementation, US_i^n user situation that include the n device preferences

3.2 Local user profile

The local profile instance represented by vector of two dimension "Users interests, document preferences" noted $< UI, UP >$

Users interests. On the same research situation the user may have multiple interests, for that we represent them with a set of vector, each one is defined by bags of user interest that scored according to the vectorial indexation space model $TFIDF$, extracted form a set of documents deemed relevant by the user. This word bags is projected on the general ontology DBpedia for semantic representation "set of concepts". The general approach of the users interests representation :

1. **Extracting a set of K relevant document** noted Kp, for every research situation, any documents wish exceeds a threshold defined during implementation is considered as relevant, and scored using the following formula:

$$Interest(Kp_i) = \alpha / N\alpha \in [1, N] \quad (2)$$

With, α is a value assigned by the user to a i document, N allowable values, Kp_i represents the i document of all documents annotated by the user, $Interest(Kp_i)$ the interest of the i, document in the Kp set.

2. **Extracting a set of term "T_i"** weighted by $tf * idf$ schema, from the Kp sub-collection of relevant document using the following formula :

$$P(T_i) = I(Kp_d) * freq_{id} * log(|Kp|/n_i) \quad (3)$$

With, $freq_{id}$ is the occurrence frequency of the term T_i in the document d, $I(Kp_d)$ the weight of document interest d in Kp collection, n_i is the number of documents in the collection Kp containing the term t_i.

3. **Projection on the general ontology DBpedia** using an approximation algorithm on the titles of the concepts. For better understanding and representation of information need, we move from keyword to a semantic concept representation. Each term T_i is compared to the *Dbpedia* concepts and represented by a corresponding "end point" concept C_i. We use *wikipidia* encyclopedia infobox to got title of concepts. and for disambiguation of the concepts we use the cosine measurement vector similarity between descriptions of the concept DC on the ontology and the word bag T_i.The concept with the lowest score will be retained using the following formula:

$$C_i = min(cos(DC_n, T_i)) \quad (4)$$

With, DC_n description of the ambiguous concepts.

4. **Calculates the degree of overlap with existing interests concepts**, with the cosine formula. If the degree of similarity does not exceed a threshold $omega$ defined on implementation, we add a new instance to this situation otherwise we merge the two vectors and we update the scores concepts using an exponential moving average. The formula :

$$P(VC_i^u)_t = \alpha P(VC_i^*) + 1 - \alpha P(VC_i^u)_{t-1} \quad (5)$$

Where,$\alpha = \frac{2}{N+1}$ a constant smoothing factor between 0 and 1, represents the degree of weighting decrease for every situations involved in the average, and N a value used to increase the accuracy of the smoothing constant.

User preference. We have presented the preferences by :

– The formats of documents, images, video, Languages

We use the frequency of each vector values in the set of user's relevant results. As preferences may depend on situations and devices, we calculate the frequency depending, on users situations using the following formula.

$$P(UP_i^j) = freq(Kp_i^d)/|US_j| \quad (6)$$

This formula represent the i preference "UP" in a situation j, which is defined by the frequency of document d that include a preference i in all documents deemed interest for a user, compared to all relevant documents for a US_j situation j.

4 Defining the source profiles

The source profiles describe the document contents and information about the source and additional information correlate with a mobile environment, we use additional parameters that correspond better to mobile devices.

4.1 Interest concepts of the source

We adopt a concepts representation based on collaborative scoring schema. Interest concepts is represented by a vector of concept we note that "VC^s" from all documents deemed relevant for all users queries $q*$. Over the same document are well noted by different users over scores of correspondent concept increase in the same source S_i, The learning process is presented by the following steps:

1. *Extraction a set of term* TS_i , weighted by $tf * idf$ schema from of all relevant documents selected by a user Kp.
2. Linking extracted vectors of words with their original sources.
3. Projection on the general ontology DBpedia ,using the same methods explained in 3.2.1
4. *Update scores of concepts* c_i *in the source* S_i with the new vector concept VC^* for each user situation by the following formula:

$$score(VC_i^s) = \begin{cases} \lambda * score(VC_i^*) + (1-\lambda) * score(VC_i^s) & \text{si} \quad i \in VC^s \\ \lambda * score(VC_i^*) & \text{si} \quad i \notin VC^s \end{cases} \quad (7)$$

Where, $\lambda = \frac{|u_d^s|}{max|u_d^s|}$ and $i \in d$and$d \in Kp$. With $|u_d^s|$: frequency of different user who select a document from the source S containing the term i, $max|u_d^s|$: max frequency of users who selected document d from the source S containing the term i. This formula increases the weight of the most recurring terms on one source "S" consulted by different user.

4.2 Source criteria

We define sources criteria by :

Documents WeightsPM. We consider the average weight of selected documents in KB, compared to the total number of documents on a source "s_i", computed as follows:

$$poids_m(S_i) = \sum_{j=0}^{|S_i|} \frac{poids(S_i^d)}{|S_i|} \quad (8)$$

With $|S_i|$ the number of document in a source S_i, $poids(S_i^d)$ weight of a document "d" by kilobyte. To normalize weight, we consider two possible states low and height calculated from the average weight of all sources, the formula described as follows:

$$P(PM_i^s) = \begin{cases} 1 & \text{si} poids_m(s_i) < \overline{PM^*)} \\ \frac{1}{2} & \text{sinon} \end{cases} \quad (9)$$

with: PM_i^s the weight of i document in the sources $s\overline{PM^*}$ represents the average of average weight of all sources.

Source adaptability. It's the ability to view the source through a wide range of devices noted RDV. We use the frequency of the responsiveness of the source relative to the number of possible resolution. we use the flowing formula to calculate the source adaptability:

$$P(RDV^s) = \frac{1}{|RL|} \sum_{i=0}^{|RL|} Adaptable(S_i^d) \quad (10)$$

With RL: the vector of all sources resolution $|RL|$: cardinality of resolution vector the Score of Adaptable is 1 if the source is responsive, otherwise 0 for a given resolution. To detect if the document is responsive and adaptable we use some indices like scrolling bar and touch-friendly sliders.

Document criteria. These criteria allow us to represent the sources according to their degree of compatibility with the documentary requirements and multimedia formats.

– The formats of documents, images, video, Languages

The compatible formats over sources, which we represent by a vector CD_i^s, each instance is weighted by the uses frequency of this format on this source throughout the trial sources scored document, the formula is as flow:

$$P(CD_i^s) = \frac{|1|}{|S^d|} * \sum_{i=0}^{|S^d|} S_i^d \quad (11)$$

For each sources we use this formula to calculate the ratio of i formats in a source,S_j^d the documents having the i format in source S and $|S^d|$ the set of all document scored in S.

5 Sources selection

In this section we present our iterative model of sources selection for a personalized access to multiple sources of information, based on user feedbacks to evaluate

previous results and improve the future requests. We consider two profile parts global "situation, device preferences" , and local "user interest, document preferences", our approach includes four steps :

5.1 Select the nearest global situation

We define a user situation by a vector $S_i =< T_i, L_i, E_i, MS_i >$ that we extend by device preferences "DP_i". finally the global situation is represented by a vector $SG_i =< T_i, L_i, E_i, MS_i, DP_i >$, the current best similar global situation SG^c is that which maximizes the sum of all dimensions similarities. We calculate the similarity between the user status and all situations S.

$$SG_j^c = MAX(\sum_{i=0}^{|SG_j|} SIM(SG_j, SG*))SG_j \in SG \quad (12)$$

$|SG_j|$ number of dimensions on a global situation, SG set of all global situation, $SIM(SG_j, SG*)$ similarity between two dimension of current global status and each instance of situations set and we have defined it as flow:

- **For temporal, geographical, event dimensions** we use a similarity of semantic affiliation in the ontologies as flow:

$$sim(C_1, C_2) = \frac{2prof(C)}{prof(C_1)+prof(C_2)} \quad (13)$$

With, $prof(C)$ the common generalizer concept.

- **For move speed dimension** two states moving or stationary considered , we calculate similarity using a flowing formula:

$$sim(MS_1, MS_2) = \begin{cases} 1 & \text{if} MS_1 = MS_2 \\ 0 & \text{sinon} \end{cases} \quad (14)$$

- **For Device preferences dimension** we use a set of vector weighted by a moving average, the similarity is calculated using cosine measure $cos(DP_1, DP_2)$.

5.2 Select the nearest local situation

The local situation represented by a user interest and preferences, the current best similar local situation depend on user interest, we use a cosine similarity measure between the local situations and query user $cos(Q, LS^n)$

5.3 Select the best sources

Similarity between interest concept of the source and concept interest of the current situation of the user.Suppose VC^u is the current interest center and VC^s the

conceptual vector of the source, we have used a cosine similarity measure, the formula is as follows:

$$Score\_ci = \frac{\sum_i^{|VC|} score(VC_i^s) score(VC_i^u)}{\sum_i^{|VC|} \sqrt{((score(VC_i^s))^2 (score(VC_i^u))^2)}} \quad (15)$$

Similarity between interest concept of the source and concept interest of the current query user.Suppose VC^u is the current interest center and VQ^s conceptual vector of the query, we use a cosine similarity measure.

The degree of consistency between the device preferences and the sources criteria.We use the devices preferences to choose the best sources that meet the device characteristics. We represent the preferences device DP by a vector { resolution,connection_speed } and the sources criteria by SC a vector of {adaptability,document_weight } To calculate the consistency between the dimensions of each vector we use the following formula:

$$Cons_pf(DP, SC) = \frac{1}{|D|} \sum_i^{|D|} \sigma_i Cons(DP_i, SC_i) \quad (16)$$

With : $|D|$ vector cardinality of DP or SC , σ_i a discrimination weight, $Cons(DP_i, SC_i)$ consistency between two dimensions DP and SC represented with a normalized cartesian distance, the formula is :

$$Cons(DP_i, SC_i) = \frac{1}{1+\sqrt{\sum_j^{|DP_i|} (DP_{ij} - SC_{ij})^2}} \quad (17)$$

The degree of consistency between the documentary user preferences and documentary requirements of the source.We use documentaries preferences to represent the interest of the user on a specific document type or language, we calculate the degree of consistency using the CU documentary criteria and sources documentary criteria CS.

$$Cons_pf(CU, CS) = \frac{1}{|D|} \sum_i^{|D|} \sigma_i Cons(CU_i, CS_i) \quad (18)$$

With: $|D|$ vector cardinality of CU or CS, σ_i a discrimination weight, $Cons(CU_i, CS_i)$ consistency between two dimensions CU and CS represented with a normalized cartesian distance.
The final sources score $SCORE(s)$ is calculated by the combination of the fourths measures:

$$SCORE(s) = \frac{1}{|measure|} \sum_i^{|measure|} \sigma_i measure_i \quad (19)$$

Where, σ_i a control variable that respect $0 < \sigma_i < 1$ and $measure = \{Score_c i, Score_q n, Cons_p f, Cons_d c\}$

Conclusion

This paper presents a contribution to the source selection in a mobile environment, we have defined user profiles and sources that incorporate dimensions to better meet the informational needs of the user as we exploit the research results to improve learning profiles.
Our research perspectives focus primarily on the experimental validation of our approach, Secondly, solving the faced problems and extending our approach with the detection of spatial and linguistic temporality and space-time fuzzy queries as well as the diversification of search results.

References

1. Pisani F., D Piotet. Comment le web change le monde : l'alchimie des multitudes, Pearson, (ISBN 978-2-7440-6261-2),188. (2008)
2. Sherman C.. Search for the invisible web. Guardian Unlimited,. étude sur les requtes et intéraction nature des requete mobile (2001)
3. Paltoglou G., Salampasis M., Satratzemi M., Modeling information sources as integrals for effective and efficient source selection,In Information Processing and Management, Volume 47, January 2011, Pages 18-36, ISSN 0306-4573.(2011)
4. Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In WWW 2013: 1201-1212.
5. Yi J., Maghoul F., and Pedersen J.. Deciphering mobile search patterns: a study of Yahoo! mobile search queries. In WWW 2008, pp. 257-266. ACM, (2008).
6. Church, K. and Smyth, B. 2009. Understanding the intent behind mobile information needs. IUI'09, February 8 -11, 2009, Sanibel Island, Florida, USA. (2009)
7. Callan J.. Distributed information retrieval. Advances in Information Retrieval, p 127–150, (2000).
8. Samet H., B. Teitler E., Adelfio M. D., and M. Lieberman D.. Adapting a map query interface for a gesturing touch screen interface. In *WWW'11 (Companion Volume)*, pages 257–260, Hyderabad, India, Mar.-Apr. (2011).
9. Divya, R.; Robin, C.R.R., "Onto-search: An ontology based personalized mobile search engine," Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on , vol., no., pp.1,4, 6-8 March (2014)
10. Callan J., Lu Z., Croft B. Searching distributed collection with inference networks. In the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, ACM-SIGIR'95, p. 21-28,1995.
11. Mishra, V.; Arya, P.; Dixit, M., "Improving Mobile Search through Location Based Context and Personalization," Communication Systems and Network Technologies (CSNT), 2012 International Conference on , vol., no., pp.392,396, 11-13 May (2012)
12. Baumgarten C., Probabilistic Information Retrieval in a Distributed Heterogeneous environment. PhD thesis, XX, (1999).
13. Hawking D., Thislewaite P. Methods for information server selection. ACM Transactions on Information Systems, 17(01):40-76, January (1999).

14. Xu J. and Croft W. B.. Cluster-based language models for distributed retrieval. In Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 254–261, (1999).
15. Fuhr N.. Resource Discovery in Distributed Digital Libraries. In In Digital Libraries '99: Advanced Methods and Technologies, Digital Collections , (1999).
16. Shokouhi M.. Central-rank-based collection selection in uncooperative distributed information retrieval. Advances in Information Retrieval, (2007).
17. Savoy J., Rasolofo Y. Recherche d'informations dans un environnement distribué. In actes "Traitement Automatique de la Langue Naturelle", TALN Lausanne, pp. 317-326. (2000)
18. Seo J. and Croft W. B.. Blog site search using resource selection. In Proceedings of ACM International Conference on Information and Knowledge Management, pages 1053-1062, (2008).
19. Hong D., Si L., Bracke P, Witt M., and Juchcinski T.. A joint probabilistic classification model for resource selection. In Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 98-105, (2010).
20. Aly R., Hiemstra D., and Demeester T.. Taily: shard selection using the tail of score distributions. In Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 673–682, (2013).
21. Bellogín A.,Gebrekirstos G. Gebremeskel , Jiyin He , Jimmy Lin , Alan Said , T. Samar† , Arjen P. de Vries† , Jeroen B. P. Vuurens, Contextual Suggestion, Federated Web Search, KBA, and Web Tracks, CWI and TU Delft at TREC, (2013).
22. Sweeney, S., and Crestani, F. Effective search results summary size and device screen size: Is there a relationship? Information Processing and Management, 42:1056-74, (2006).
23. Schusteritsch, R., Rao, S., and Rodden, K.. Mobile search with text messages: designing the user experience for Google SMS. In CHI '05 Extended Abstracts on Human Factors in Computing Systems CHI '05. ACM Press, New York, NY, 1777-1780. (2005)
24. Kamvar M Query suggestions for mobile search: under- stationary usage patterns. In: Proceedings of CHI, pp 1013–1016 (2008)
25. Kamvar, M. and Baluja, S. The Role of context in Query Input: Using contextual signals to complete queries on mobile devices. In Mobile HCI Proceedings, 121-128. (2007)
26. Bouidghaghen, O.; Tamine, L.; Boughanem, M., "Context-Aware User's Interests for Personalizing Mobile Search," Mobile Data Management (MDM), 2011 12th IEEE International Conference on , vol.1, no., pp.129,134, 6-9 June 2011
27. Iwata M., Miyamoto H., Hara T., Komaki D., Shimatani K., Mashita T., Kiyokawa K., Uemukai T., Hattori G., Nishio S., and Takemura H.. 2013. A content search system considering the activity and context of a mobile user. Personal Ubiquitous Comput. 17, 5 (June 2013), 1035-1050.

Image Segmentation by Image Analogies

Asma Bellili and Slimane Larabi

LRIA, USTHB,
USTHB, BP 32 El Alia, Bab Ezzouar Algiers, Algeria

bellili-asma@hotmail.fr, slarabi@usthb.dz

Abstract. In this paper we propose a new technique for image segmentation based on contour detection using image analogies principle. A set of artificial patterns are used to locate contours of any query image. Each pattern allow the location of contours corresponding to specific intensity variation. Boundaries are extracted based on the properties of located contours. In addition, elementary regions derived from the motion of contours in images are located and combined jointly with the boundaries for image segmentation. Experiments are conducted and the obtained results are presented and discussed.

Keywords: Image Segmentation, Analogies, Contour Detection, Multi- Scale .

1 Introduction

Image segmentation is a preprocessing step whose goal is to express an image in meaningful way and to divide it into spacial regions having some common characteristics. This task which change the representation of an image into something that is easier to analyse is a fundamental process in many computer vision ap- plications. There are many image segmentation methods proposed in the literature. Many states of the art have been done and published [6] [14] [15] . A review of the literature in image segmentation indicates that natural images segmentation algorithms can be divided into two categories : Region based and Edge-based approaches. Region based approaches aim to regroup pixels having similar attributes, and the edges-based methods aim to separate regions having dissimilar attributes. This problem remains an attractive topic for two reasons: the first one is that the results of proposed techniques are still far from what can do the human. The second one is that the segmentation is a critical step for all applications. Image analogy is a new technique for image processing by example which consist in two steps: - The first one consist on designating two images (A,A0) such that A0 is the transformation of A applying a filter. - Assuming that the transformation between (A,A0) is learned, the second step consist to apply to any given image B the same transformation (A : A0) giving the image B0 [7] [10].

Image analogies has been largely used in different applications such as super resolution [8], texture [2] [3] [5], curves synthesis [11], image colorization, image enhancement and artistic filters [16], [17]. An advantage of image analogies technique is the possibility to learn very complex and non linear image filters such as artistic filters in witch various drawing and painting styles are synthesized based on scanned real world examples [10].

Few works have been devoted for the use of image analogies in image processing. A method for supervised segmentation of medical images is proposed by Lackey and Colagrosso [12] applying directly and naively the algorithm of Hertzmann [10]. This method is applied only to find by analogies the same coloured regions in medical images as those processed by the expert. We notice here that the application of naive way to image segmentation like is made in [12] requires numerous pairs of learned images (A,A0) where A is initial image and A0 is the segmented image. This is due to the requirement of all lighting conditions in learned images. S. Larabi and N. M. Robertson proposed a method based on the learning of the expertise of hand draw contours to locate outlines of a query image in the same way that is done for the reference [13]. The result of their approach was a set of images which contains several contours, each one is the result of using of artificial pattern instead of hand drawn contours. In [4], authors proposed a method based on these contours in order to define and locate the outlines of objects. In this work, we propose a method to extract boundaries of objects and then to segment image in regions. The next section is devoted to a brief review of contour detection by image analogies and the inferred properties. Our method is presented in section 3 followed by the results of experiments conducted on Weizmann data set [1] presented section 4.

2 Contour detection by image analogies and properties: A brief review

The idea is to start from a pair of images (A,A0) , A is an initial image and the second one A0, identical to A, in addition, contours are hand drawn. The aim was to localize the contours of any other query image B using the expertise learned from the pair (A,A0) (see figure 1)[13]. A set of artificial patterns are proposed instead of hand drawn contours in images as learning images (see figure 2). The use of these patterns allow locating fourteen images of contours. Each one is obtained by the corresponding pattern (A,A0). In [4], authors demonstrated that contours in the 14 images of contours lo- cated by image analogies technique are moving from one image to another and are more steady around the boundaries of regions. However, for others parts of the image, they are moving fastly (see figure of table 1).

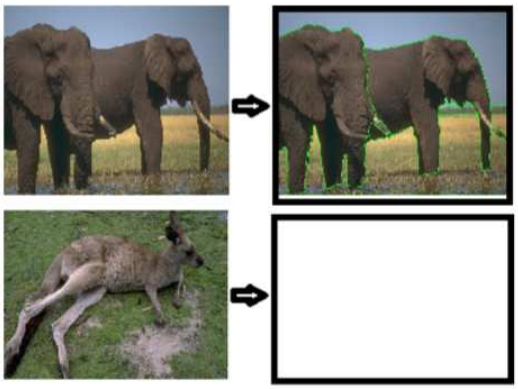


Fig. 1 Contour detection: The main idea

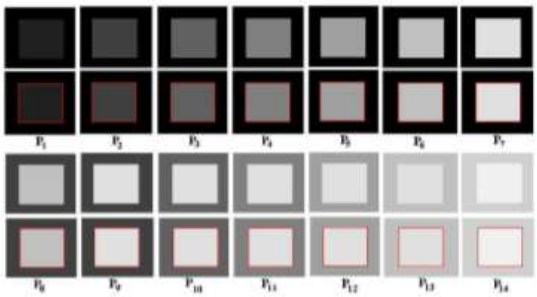


Fig. 2. Artificial patterns used as training images

Table.1. Contours located using a selection of the 14 artificial patterns

3 Image segmentation

3.1 Energy map of pixels

The stability of a contour is measured from its motion in all images of contours. More the motion of the contour is slow, more it will be considered as the region boundaries [4] (see figure 3).

In this section, we propose firstly an algorithm for measuring the stability of a pixel among all images of contours. A map of energy is created from the images of contours and used to locate regions. Also, depending on the energies of pixels, multi-scale segmentation is presented.

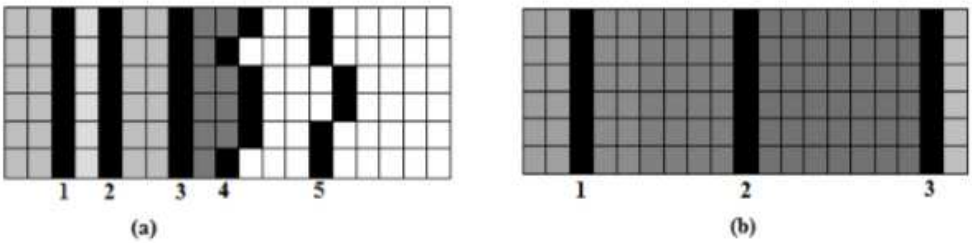


Fig. 3. (a)Slow motion of contours located using five successive patterns and reported in the same part in image, (b) Fast motion of contours located using three successive patterns and reported in the same part in image

Let pki,j be the pixel (i,j) in image IMk of contours obtained using the pairs of patterns Pk,k = 1..14. There are 14 images IMk and in each one the pixel pki,j may be or not a contour pixel. Our aim is to compute an energy map of pixels starting from the set of images of contours. Each pixel p(i,j) will be associated a value measuring its appertaining to a border of region or to inside of region. As explained above, around the pixel p(i,j) in the image, pixels of contours of all images IMk,k = 1..14 are moving from the darkest part to the clearest one (see figure 4).

Fig. 4. Appearance of the same pixel (i,j) for different images IMk of contours

To evaluate the evolution of the contour around p(i,j) in all images of con- tours IMk,k = 1..14, we consider an area of ((2N+1)×(2N+1)) pixels. For each pixel p, and for each IMk, we search the nearest pixels of contours q following the n directions in the defined area. Four directions are considered: Horizontal, vertical and two diagonals directions. The energy of the pixel p is computed using the distances dk of the nearest contour pixel qk i,j to p in the image IMk. The energy E of p is defined by the following equation :

$$E = \sum_{k=1}^{k=14} 2^{N-d_k}$$

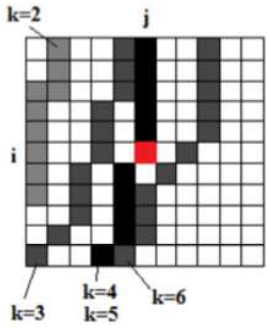


Fig. 5. Search area of contour pixels for all 14 images. Notice the presence of five contours located using five successive patterns k = 2..6, two of them pass by p(i,j)

More there are pixels q nearest to p(i,j) found in successive images IMk,k = 1..14, more the energy E is highest. However, when there are few (one or zero) pixels q found, this implies that the energy E is very low. For example, E = 2 is associated to p(i,j) in case where there is only one pixel contour located at distance (N − 1) from the central pixel (i,j). However, when there are five contours pixels located (such that illustrated by figure 5), corresponding to the distances N − 1, 1, 0, 0, 1 (moving from the left to right), the energy E is then equal to E = 21 + 2N−1 + 2N + 2N + 2N−1, for N = 5, E = 98. This energy is then synonymous of the appertaining of the pixel p(i,j) to the inside or to the border of region. The values of energy computed have similar significance like the result of merging of all images of contours giving largest contours at borders of regions and thinnest elsewhere (see figure 6).

Fig. 6. Result of merging all 14 images of contours

3.2 From map of energies to multi-scale segmentation

The different values of energy associated to each pixel define the map of energies end noted Im. We used this map to locate the boundaries of regions at different scales: (high, intermediates and low) which correspond to the four intervals of energy: [46,56[, [56,64[, [64,128[, [128,∞]. For each value of energy, there are pixels of boundaries which are selected, where strong boundaries are associated to high values of energy. Figures of table 2 illustrate for each image the located boundaries at low, intermediate and high scale.

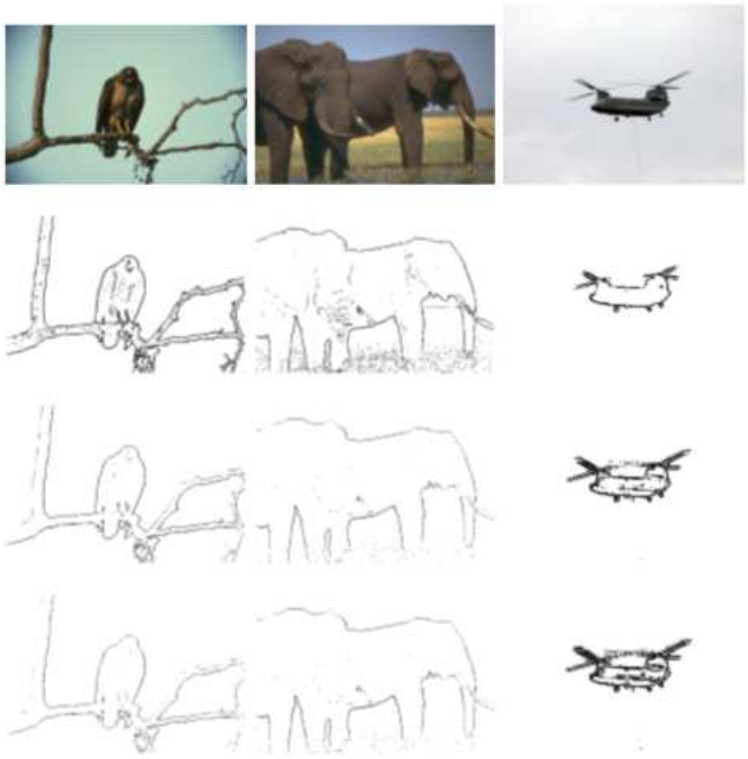


Table 2. For some images of Weizmann data set, boundaries located at low (second row), intermediate (third row) and high (fourth row) scales

Once the boundaries are located, an additional processing is required in order to achieve the segmentation. Elementary regions defined as the areas between contours located using two successive patterns are firstly located and combined jointly with the located boundaries in order to locate regions in image. Figure 7 illustrates these elementary regions where 14 colors are associated to elementary regions obtained using the 14 artificial patterns.

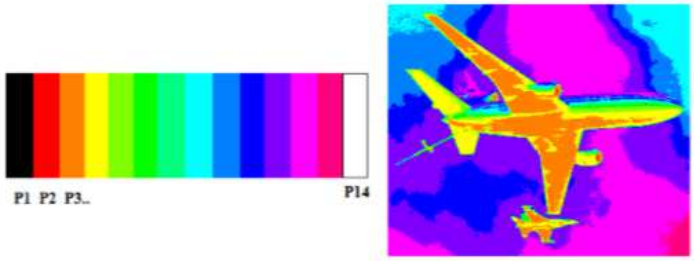


Fig. 7. At the left: the different colours associated to elementary regions located using each one of the 14 patterns. At the right, the result obtained for image from Weizmann data set.

4 Results

In this section we present results obtained by applying our method on Weizmann data set. Figures of table 3 show the boundaries detection of images at high scale, the elementary regions located using the 14 patterns and the result of fusion with computed boundaries. Compared to the ground truth data, we achieved a good values of recall and precision (see figure 8).

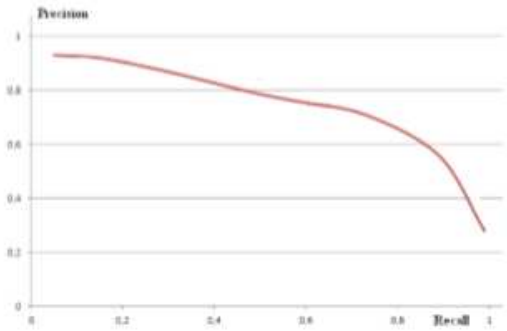


Fig. 8. Evaluation of the obtained results on Weizmann data set

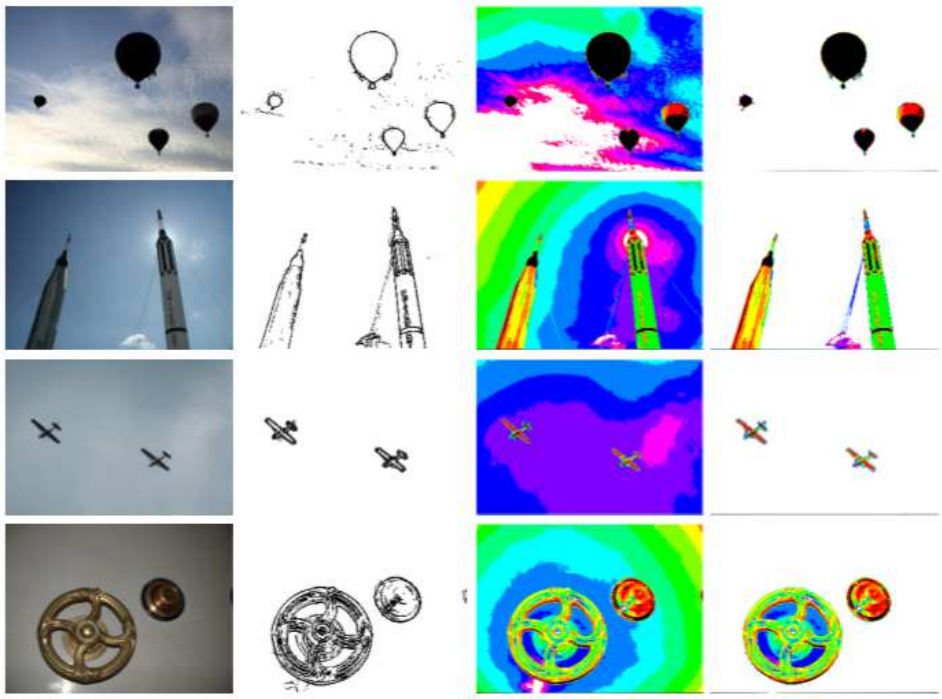


Table 3. For some images of Weizmann data set, boundaries located at high scale, the located elementary regions, the result of fusion process.

References

1. Sharon A and Meirav G and Ronen B and Achi B, Image Segmen- tation by Probabilistic Bottom-Up Aggregation and Cue Integration, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June, 2007.
2. Ashikhmin, M. (2001). Synthesizing natural textures. In Proceedings of 2001 ACM Symposium on Interactive 3D Graphics.
3. Ashikhmin M., Fast texture transfer. In IEEE Computer Graphics and Applications 23(4): pp. 38-43, 2003.
4. Bellili A., Larabi S., Robertson N. M., Outlines of objects detection by analogy , In In the LNCS proceedings of 15th on Computer Analysis of Images and Patterns (CAIP'2013), pp. 385-392, August 27-29, York, UK

5. Bhat P., S. I. and Turk, G. (2004). Geometric texture synthesis by example. In Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing, Nice.
6. Cheng, H.D., J. X. H. S. Y. and Wang, J. (2001). Color image segmentation: advances and prospects. In Pattern Recognition.
7. Cheng L., S. V. and Zhang, X. (2008). Consistent image analogies using semi- supervised learning. In IEEE Conference on Computer Vision and Pattern Recog- nition (CVPR), Anchorage.
8. Freeman W. T., Pasztor E. C., Carmichael O. T., Learning Low-Level Vision, In International Journal of Computer Vision 40(1), pp. 25-47, 2000
9. Hertzmann A., Jacobs C. E., N. O. B. C. and Salesin, D. H. (2001). Image analo- gies. In Proceedings of the 28th annual ACM conference on Computer graphics and interactive techniques, New York.
10. Hertzmann A., Jacobs C. E., Oliver N., Curless B., S. M. Seitz, Image analogies. In SIGGRAPH Conference Proceedings. 2001. pp. 327-340
11. Hertzmann A., Oliver N., B. C. S. M. S. (2002). Curve analogies. In EGRW '02 Proceedings of the 13th Eurographics workshop on Rendering, Switzerland.
12. Lackey J. B. and Colagrosso M. D., Supervised segmentation of visible human data with image analogies. In Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications, 2004.
13. Larabi S., Robertson N. M. (2012). Contour detection by images analogies. In Proceedings of 8th International Symposium on Visual Computing, Crete,Greece.
14. Nikhil R Pal, S. K. P. (1993). A review on image segmentation techniques. In Pattern Recognition.
15. Haralick, R. M. and Shapiro (1985). Image segmentation techniques. In Computer Vision, Graphics and Image Processing.
16. Sykora D., Burianek Zara J., J., Unsupervised colorization of black-and-white cartoons. In Proceedings of the 3rd Int. Symp. Non-photorealistic Animation and Rendering, pp. 121–127, 2004.
17. Wang G., Wong T. and Heng P., Deringing cartoons by image analogies. In ACM Transactions on Graphics, Vol. 25, No. 4, October 2006, pp. 1360-1379.

Texture Analysis and Matching

Izem Hamouchene and Saliha Aouat

LRIA, USTHB,
USTHB, BP 32 El Alia, Bab Ezzouar Algiers, Algeria

{ihamouchene, saouat}@usthb.dz

Abstract. Texture matching is an important field of image processing at pattern recognition. This task consists of finding a particular texture into an image that contains several texture. Matching is a difficult problem to solve and remains an attractive area for research. However, several and recent approach have been proposed for texture matching. Inspired by the last studies, we proposed a new architecture for texture matching. This architecture is called the Dynamic Decomposition Architecture (DDA). The main idea of this method is to fix a converging point α. After that, decomposing the image in the direction of this point. In this work, a main window starting from the converging point α to the bottom-right corner of the image is considered. After that, the size of the main window is reduced and other windows with the same size of the main window are generated. The Local Binary Pattern (LBP) technique, which is invariant to monotonic grey level changes, is applied of the ex-tracted windows to describe the texture. Synthetic images are used in the experimentations using the proposed architecture and some obtained results are illustrate.

Keywords: Texture matching, texture analysis, features extraction, texture segmentation, decomposition architecture, LBP.

1 Introduction

The automatic content based image processing have become a dynamic research area since 1990. Segmentation, matching and indexing are the most important tasks of image processing. These tasks analyze automatically the image based on visual contents. Image processing is subdivided into three categories: Color, Shape and Texture. Texture is present in most of real life objects in nature. This makes it fundamental and essential to analyze images. The last category (Texture) is the studied category in this work. Texture can be subdivided into coarse, micro, macro, regular, periodic, aperiodic, random and stochastic type [1].

Texture analyses has been introduced by Haralick [2]. Different approaches have been presented : structural, statistical and transformed. The transformed approach transform the original image into another domain such as frequency domain (Gabor [3, 4] or wavelets).

These approaches have been used in different applications, especially in texture matching; the principle is to have a texture recognized in an image which contains several different textures.

Texture matching consists of recognizing and segmenting a particular texture on an image that contain several texture. Most proposed methods to solve this problem (matching) requires adjustment of parameters. This adjustment are crucial and more parameters are adjusted, the better the results will be. These methods use a decomposition architecture to localize the researched texture. The goal of this architecture is to decompose the image into blocks. After that, fea-tures are extracted from each block. Finally, a similarity measure are calculated between each block and the researched texture in order to decide if this block is accepted or not. One most disadvantage of this classical architecture is the fixed size of the blocks. Indeed, if the size of one block is too big, one block may contain different texture. In the other hand, if the size on the blocks is too small, the researched texture may not be recognized. The size of the decomposition is crucial and affect the quality of the recognition system. However, the problem of the decomposition size is a difficult problem to resolve.

In this study, we have proposed a new decomposition architecture for texture analysis. The focus of this work is on the decomposition scale in order to solve the fixed size of the generated blocks. For the feature extraction step, any feature extraction method can be applied. In this work, we have chosen the LBP (Local Binary Pattern) method which is an adapted method for texture analysis and it is also invariant to rotation [5].

This paper is organized as follows: The next section we explain the Local Binary Pattern (LBP). In section 3 we present our proposed decomposition architecture. Section 4 illustrates experimental results using the proposed system and the last section conclude the paper.

2 The LBP method

The Local Binary Pattern (LBP) is a simple and efficient feature extraction method [5]. This method was presented by Ojala and Pietik¨ainen [6, 7]. This method unify the statistical and structural approaches. The LBP method is a monotonic gray-scale transformation invariance and rotation invariance [12, 13]. This method has been applied in various and recent works [8, 9, 10, 11].

The idea of the LBP is to thresholding the 3x3 neighborhood of each pixel. Then, consider the result as a binary number. Thus, each pixel of the analysis window is thresholder by the value of the central pixel. This neighbor is encoded by the value 1 if its value is greater than the central pixel and 0 otherwise. The obtained binary code is converted into a decimal number and represent the new value of the central pixel on the

output LBP image. This process is illustrated in Figure 1.

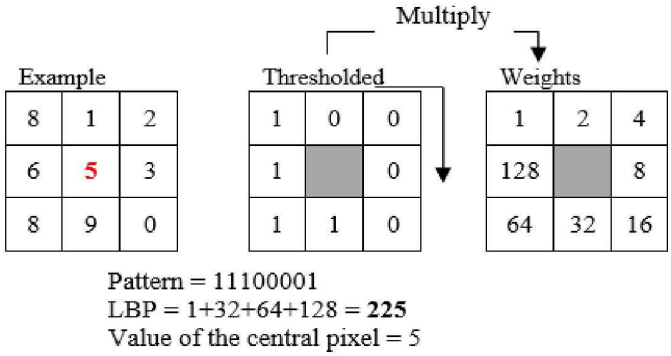


Fig. 1. Calculation of the binary pattern

Figure 1 illustrates the calculation of the LBP number. The analysis window (3x3) are extracted from the image. Each neighbor is encoded by the value 1 or 0. Finally, the binary code is converted into a decimal number (225) that represent the value of the central pixel on the LBP image.

In order to describe the output LBP image, a histogram (of the 256 possible patterns) is created to collect up the occurrences of each pattern [14]. This histogram ,which represents the statistical occurrence of the local-micro patterns, is normalized and considered as a feature descriptor of the texture. This process is illustrated in Figure 2.

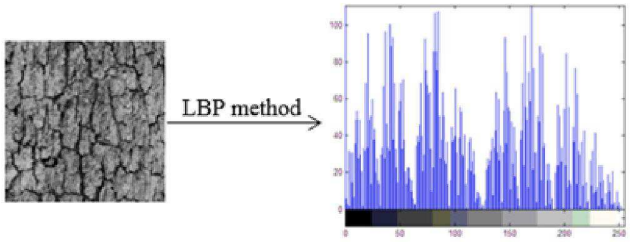


Fig. 2. Feature extraction using LBP method

3 Proposed decomposition method for texture analysis

In this section, we will explain the proposed decomposition architecture. Some examples will be illustrated to explain the proposed process.

The proposed decomposing architecture follows a series of steps. First, a converging point a is fixed. This point represent the direction of analyze. After that, a Main analysis Window (MW) is considered from a to the button right of the image. The main window is extracted and its LBP histogram is calculated. Then, the similarity measure between the LBP histogram of the MW and the researched texture is calculated. If this measure is above the threshold, this window is considered as pertinent. After that, the size of the MW is reduced and as many Windows with the same size of the MW are extracted. This process is repeated until the MW's size reach un minimum size. The algorithm of the proposed Dynamic Decomposition Architecture (DDA) is illustrated in algorithm 1.

Algorithm 1 Square dynamic decomposition system

1. Choose the converging point α.
2. Consider the main window MW (from α to the button right of the image)
3. Generate as many windows as possible, with the same size of MW
4. **for** each window **do**

   (a) Apply the LBP method
   (b) Extract the LBP histogram
   (c) Calculate the similarity Sim between the extracted window's histogram and the sought texture histogram
   if Sim is above the threshold **then**
   Save the window's coordinates and its histogram in a vector V
   end if

   end for
5. Reduce the size of MW by the distance d

if the size of MW is below the minimum size μ **then**
Color the saved windows and terminate the process else
Goto step 3
end if

The point a (0, 0) is fixed into the top left of the image (Figure 3). The similarity measure between the MW and the researched texture is calculated using formula 1.

$$Sim(His\backslash,His2) = E™=imin\{His\backslash[i\backslash, His2[i\backslash)} \quad (1)$$

Where his is the window's histogram and hist2 is the sought texture's histogram.If this similarity (Sim) is above the threshold, this window is considered as pertinent. The coordinates and histogram of this window are saved in one vector Vi=[X, Y, Height, Width, Histogram]. This process is illustrated in Figure 3.

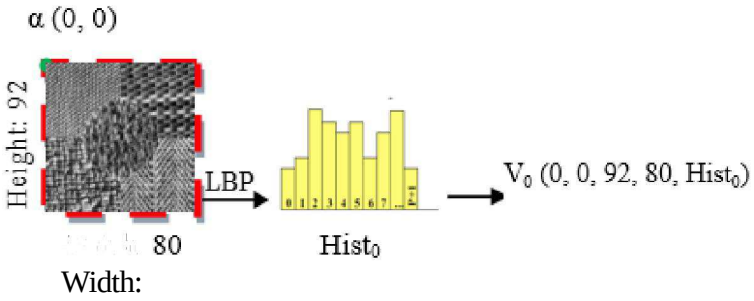


Fig. 3. Feature Extraction from the main window

After that, the size of the MW is reduced. This process is repeated and many windows with the same size of MW are generated. This process is illustrated in Figure 4.

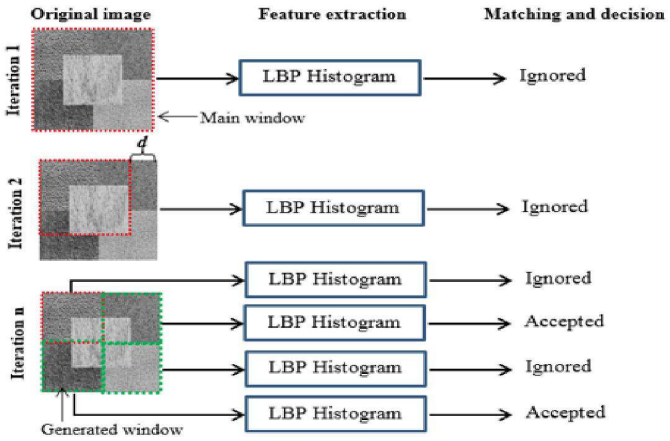


Fig. 4. Iterations of the decomposition process

Figure 4 illustrates the different extracted window in different iterations. Each iteration, the main window (red) and other windows (green) are generated.

This process is terminated when the main window's size reaches a minimum size μ, it means the height or the width of the main window is less then μ.

Graphic illustration of the proposed dynamic decomposition architecture is illustrated in Figure 5.

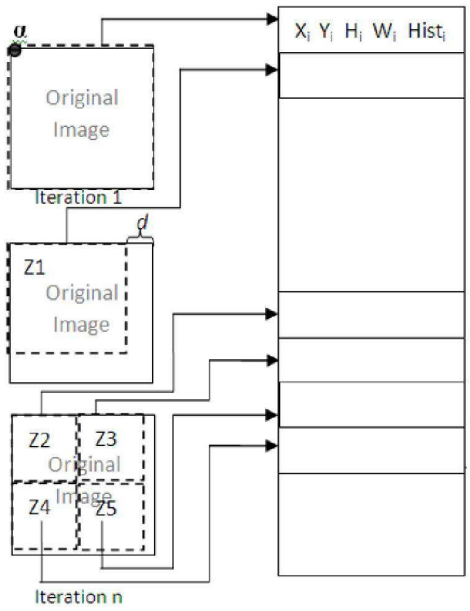


Fig. 5. Illustration of the decomposition process

Figure 5 summarize the proposed architecture. For each iteration, the main window and other window with the same size of the MW are generated. A similarity measure is applied and only the pertinent windows are saved. Finally, the stored windows are plotted.

The most important advantage of this architecture is to analyze windows with different size. Thus, a rich and various set of windows are extracted and compared with the researched texture. Another improvement can be applied is to use a different converging points. Therefore, different configurations can be studied. Using this architecture, there are no constraints in the feature extraction step, so any features

extraction method can be used to describe the texture. Thus, the more adapted method to the researched texture should be used, which represents another advantage of our technique.

4 Performance and evaluation

In order to evaluate the proposed architecture, we applied the proposed system on six synthetic images that are composed of different textures (see Figure 6).

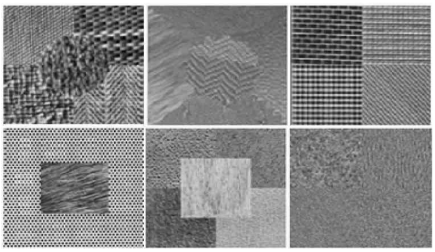


Fig. 6. Synthesis images used in the experiments

In our study, the distance of the recusing size of the MW is fixed to 10 pixels. The converging point was assigned to (0,0). And the minimum size is limited to 5 pixels. The threshold was fixed to 0.9. The evaluation process was as follow: First, one window is extracted from the test image. This window contain one texture (researched texture). After that, the proposed system is applied to recognize the south texture. So, the output image contains the most similar texture to the researched texture (the most similar texture is colored with red). This process is illustrated in figure 7.

The figure 7 illustrates an application example. The query is extracted from the test image and detected as shown in the resulted images. We can notice that this system gives a better recognition when the researched textures have a square shape.

In order to compare the proposed architecture with the classical one, we applied the classical decomposition architecture. The test image is decomposed into 32x32 blocks. The LBP method is applied on each block to extract the feature. The obtained results are illustrated in Figure 8.

We can notice that this architecture localize approximately the sought texture but some textures are not well recognized, some images have the problem of border detection, other scale problem (fixed 32 pixels may not be the best scale).

We can notice that the obtained results using the proposed system (Figure 9) are remarkably better than the classical method. In fact, the weakness of the classical

architecture (border detection, fixed scale) are improved using the proposed architecture. This This is due to the dynamic size of the considered windows This allows us to analyze the image with different scales. This represents the main advantage of the proposed architecture.

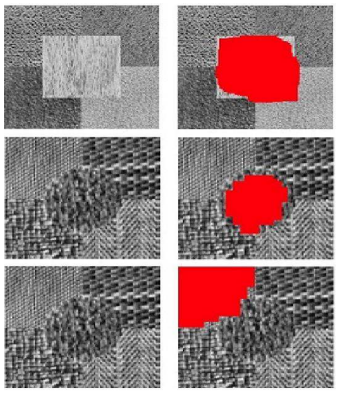


Fig. 7. Illustration of the texture matching results.

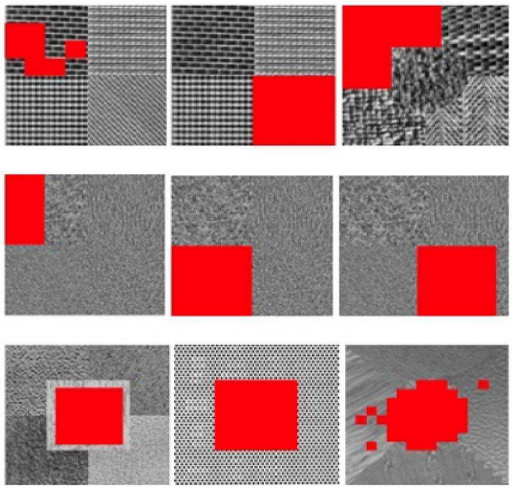


Fig. 8. Experimental results using the classical architecture.

"

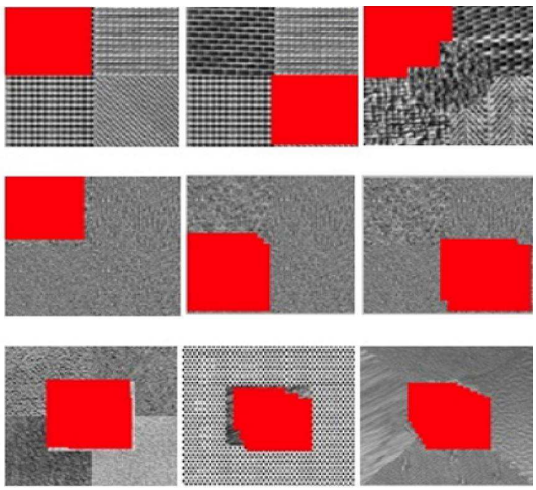


Fig. 9. Experimental results using the dynamic decomposition architecture.

5 Conclusion

In this study, we presented a new decomposition architecture for texture analysis. This architecture generates windows with different size unlike the traditional method. The Local Binary Pattern (LBP) is applied to describe the texture of each extracted window. Finally, a similarity measure is calculated between each LBP histogram extracted from windows and the researched texture. If the similarity measure is above the threshold, the window is considered as pertinent and colored by red. In the experimental part, we applied the proposed system on synthetic test images. The proposed system has recognized better the researched texture compared to the traditional architecture. This is due to the different size of the extracted windows. In addition, there are no constraints in the feature extraction step, so any feature extraction method can be used to describe the texture. This architecture can be also used to segmentation or indexing images. In future studies, we will study the application of different shape of the analysis window. We will also study robustness of the proposed architecture applied on real textured images.

References

1. Richards, W and A Polit, "Texture matching", Kybernatic, 16, pp. 155 – 162, 1974.

2. Harlick, R., "Statistical ans structural approaches to texture", Proc. Of IEEE, vol. 67, no. 5, pp. 786-804, May 1979.
3. Bovik, A.C.; Clark, M.; Geisler, W.S.; , "Multichannel texture analysis using local-ized spatial filters," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.12, no.1, pp.55-73, Jan 1990.
4. Levesque, V., "Texture segmentation using Gabor filters", Center For Intelligent Machines, Mc Gill university December 2000.
5. Ojala, T., Pietik¨ainen, M. and M¨aenp¨a¨a, T.: "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns". IEEE Transactions on Pattern Analysis and Machine Intelligence 24 971 – 987, 2002.
6. Ojala, T., Pietik¨ainen, M. and Harwood, D., A, "Comparative Study of Texture Measures with Classification Based on Feature Distributions", Pattern Recognition 19(3):51-59, 1996.
7. T. Ojala and M. Pietik¨ainen, "Unsupervised Texture Segmentation Using Feature Distributions", Pattern Recognition, vol. 32, pp. 477-486, 1999.
8. Baohua, Y., Yuan, H. and Jiuliang, C. "Combining Local Binary Pattern and Local Phase Quantization for Face Recognition ", Biometrics and Security Technologies (ISBAST), pp. 51-53, March 2012.
9. Xueming, Q., Xian-Sheng, H. and Ping C., Liangjun, K. "An effective local binary patterns texture descriptor with pyramid representation" ,Pattern Recognition, Vol. 44, Issues 10-11, pp. 2502–2515, November 2011.
10. Zhenhua Guo; Lei Zhang; Zhang, D.; , "A Completed Modeling of Local Binary Pattern Operator for Texture Classification," Image Processing, IEEE Transactions on , vol.19, no.6, pp.1657-1663, June 2010.
11. Kellokumpu, V., Zhao, G., and Pietik¨ainen, M. "Recognition of human actions using texture descriptors ",Machine Vision and Applications, 22(5):767-780, 2011.
12. Cuiyu, S., Fengjie, Y. and Peijun, L. "Rotation Invariant Texture Measured by Local Binary Pattern for Remote Sensing Image Classification", Education Technology and Computer Science (ETCS), vol.3, pp.3-6 , 2010.
13. Zhenhua, G., Lei, Z. and David, Z. "Rotation invariant texture classification using LBP variance (LBPV) with global matching",Pattern Recognition Vol. 43, Issue. 3,pp. 706–719, March 2010.
14. M¨aenp¨a¨a Topi, Ojala Timo, Pietik¨ainen Matti and Soriano Maricor, "Robust Texture Classification by Subsets of Local Binary Patterns," icpr, vol. 3, pp.3947, 15th International Conference on Pattern Recognition (ICPR'00) - Volume 3, 2000.