Recommending Competent Users from Semantic Traces using a Bayes Classifier

Ning Wang, Marie-Hélène Abel, Jean-Paul Barthès
UMR CNRS 7253 HEUDIASYC
Sorbonne universités, UTC
Compiègne, France
{ning.wang, marie-helene.abel, barthes}@utc.fr

Elsa Negre
UMR CNRS 7243 LAMSADE
Université Paris Dauphine
Paris, France
elsa.negre@dauphine.fr

***Abstract*— When users collaborate, they leave traces in some way or another. These traces in return offer a clue whether a user is competent enough on a subject. This helps further collaboration because knowing the specialization of users helps to distribute tasks reasonably. In this article, we propose a semantic model of traces and analyze classified traces using a Bayes classifier. We exploit the results to offer recommendation on competent users accordingly.**

Keywords: traces of interaction; competency; semantic modeling; recommender system; Bayes classifier

I. INTRODUCTION

Collaboration is a source of power for human society development and progress [1]. It requires transactive memory including two parts: (a) a combination of individual knowledge and (b) interprersonal awareness of others' knowledge [2]. Previous research has found a positive relationship between transactive memory system development and group performance [3]. Thanks to information technology, transactive memory system can be applied in informational environment where users meet their goals by taking different actions more easily. We are interested in the effects of actions as well as in the actions themselves. A set of actions, step by step, is defined as a trace [4]. Under modelling and analysis, traces in return help indicating the competency of an individual [5]. Based on the information exploited from the traces, we improve collaboration focusing on the reuse of traces for different purposes like decision aid and recommendation [6].

This paper proposes a prototype that models, records and analyzes users' traces. It allows recommending people with more expertise on a certain subject. To achieve this objective, the following tasks are needed: (i) proposition of a semantic structure to record traces; (ii) proposition of a model of competency; (iii) evaluation traces using a Bayes classifier and semantic distance and (iv) proposition of recommendations accordingly.

In Section 2 we identify various limitations of the current studies on recommendation. In Section 3 we propose a recommender system for exploiting traces. In Section 4 we illustrate our method by an example. Section 5 gives conclusions and points to directions for future works.

II. RELATED WORK

The interest of a recommender system is justified by the need to manage the growing amount of information [7]. Recently various articles have been published about exploiting the traces with the help of semantics. Chen et al [8] present a mechanism for personalized knowledge search and recommendation adapting a suitable domain ontology according to the previous browsing and reading behavior of users. Sahay et al [9] propose a novel conversational search and recommendation system that involves finding relevant information based on social interactions and feedback. Breese et al [10], Condliff et al [11] and Pennock et al [12] all tried to provide recommender using probabilistic modeling. But none of these studies focused on combining action model with a probabilistic method for the recommendation of users' competency.

Our previous work tried to give a solution by TF-IDF [13], but it lacks capability when the number of features deciding a recommendation is large. Inspired by Ghazandfar & Adam [14], Schein et al [15] and Melville et al [16], in the following we give recommendation based using a Bayes classifier.

III. OUR APPROACH

We orchestrate a model of actions, a model of competency and a Bayes classifier to make recommendations about users' competency. Fig. 1 shows the structure of recommender system. Firstly, users' actions are collected and modelled from an interactive platform. After being sifted by the filter of classification, we obtain classified traces, which allows a preliminary presentation back to the users. Alternatively, we apply an algorithm to calculate an index indicating the correlation between the classified traces of a certain user and a given subject. These values can lead to useful information that are presented as personalized recommendations, either to a group defined as a set of users of the platform, or to an individual user.

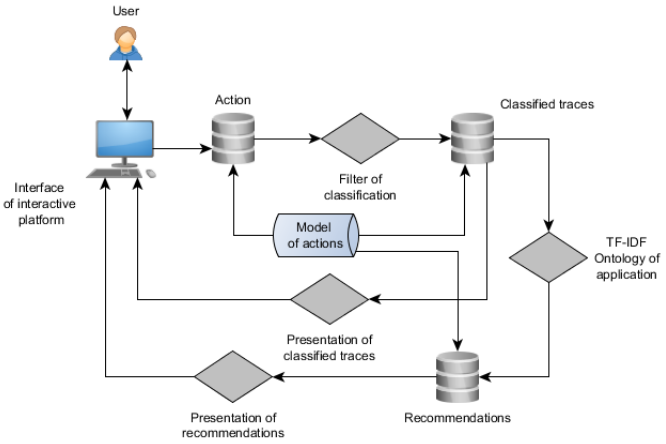

Figure 1. The structure of our proposed recommender system for the exploitation of semantic traces.

A. *Model of Actions*

We define the principal concepts as follows:

- Action: an interaction or an act performed by a user in a collaborative environment, e.g. sending a document to other users.
- Classified trace: a set of actions that were performed by a user in the informational environment classified according to the model of traces [17].
- Set of traces: an ordered set of classified traces.

According to our definition, an action is the basic element forming a trace. Regarded as an important resource for our recommender system, we introduce the Resource Description Framework (RDF) to model actions [18]. RDF is used as a general formalism for conceptual description or modelling of information that is implemented in web resources. Fig. 2 shows the basic structure in the RDF schema of our model. An ellipse represents a class of resources and a rectangle represents an object property. For example, a person has the object property "has_id_person" and the range of this property is a class called "id".

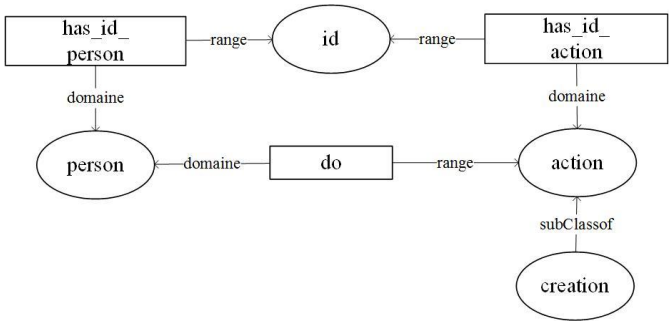

Figure 2. Basic structure in RDFS graph presenting an action.

This model of actions has two main advantages compared to a traditional form of history or log of users:

- Actions are presented in a labeled, directed multi-graph. In our model actions are represented as resources in the RDF schema and they are connected by properties. This allows a better structure of storage and usage of actions. For example, a person "Ala" chats with "Ning." This action can be represented by an RDF instance showed in the lower part of Fig. 3 where "Ala" and "Ning" are two instances of the class of resources "person." "Chat_1" is an instance of the class of resources "conversation" which is linked to the action "creation."
- Normally, different types of actions have different importance. For example, creating elements of a Wiki is more important than consulting it. In our model actions are classified using three classes: creation, consultation and addition, which enables to treat different types of actions more efficiently. We define the importance of creation, addition and consultation respectively as "high", "medium" and "low."

B. *Model of Competency*

Recently, many researchers have focused on modelling competency. Molina and Flores [19] define a "Core Competency" in the manufacturing clusters, including 4 generic and comprehensive components, namely: products, processes, skills, and task service. Müller [20] proposes "Competence cells" for the competency cell-based networks in which the main components are "resources" and "fulfilled task or executed function". Boucher et al [21] present a "s-a-r-C model" of competency consisting of "Professional Situations," "Actor" and "Resource." These models share two components, namely "resources" (including "human resources" or "physical resources") and "activity" (also called "process," "production skill" or "task").

The success of an activity requires actions on relevant concepts. For example, if we want to create a website, collaborators should put their knowledge about different concepts like "PHP," "Javascript" and "HTML" into effect. During the activity, their actions are recorded, e.g., creating a manual or sharing a technical article. Our system aims at analyzing these actions and evaluate collaborators' competencies on different subjects, so that when the next time arrives and a certain activity needs an expert on a certain concept, we can recommend a collaborator. As we are interested in the management of knowledge, resource is represented mainly in the form of knowledge. Thus we propose an "action-knowledge model" that integrates the merits in the models above and covers the strength of our system as shown in Fig. 3.

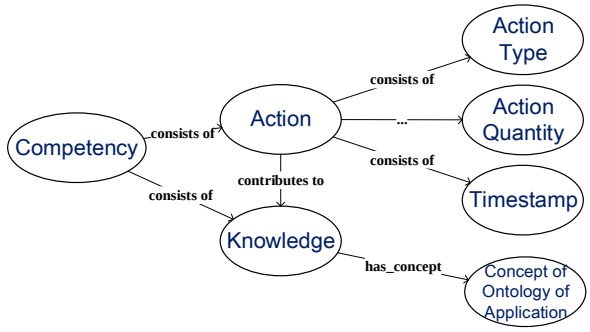

Figure 3. Generic a-k model of competency

The detailed definition of these components is the following:

- *Action* is how a user applies the knowledge. Action also helps accumulating the knowledge of a user. For example, if a user consults many files about "Java," it's reasonable to assume that his/her knowledge grows.
- *Action Type* describes different types of actions. Some types of actions directly contribute to competency, for example answering questions from other users or creating a Wiki about this concept. Such actions indicate that the user tends to be more competent about what he/she applies. Meanwhile other actions only contribute to the knowledge of the concept such as reading a paper about it. In III.A action types are described by semantic model of actions.
- *Action Quantity* records both a user and his collaborators' intensity of efforts on this activity.
- *Timestamp* records the time when an action took place.
- *Knowledge* is what a user applies during an action.
- *Concept of Ontology of Application* describes the nature of a user's action. It is the semantic description of knowledge.

With this model of competency we merge our methods in this section to evaluate users' competencies.

C. Application of a Bayes Classifier

Previously, we focused on analyzing traces using TF-IDF [13]. Because a trace is composed of actions on a set of concepts, we need a method that better handles multi-dimension factors. The Naïve Bayes classifier is based on Bayes theorem with a strong (Naive) independence assumption, and is suitable for the cases having high input dimensions as indicated by Ghazanfar and Adam [14]. In the following, we elaborate on adapting the method to our purposes.

Naïve Bayes is a conditional probability model. Given a problem instance to be classified, represented by a vector of features $F=(F_1,...,F_n)$, we tend to calculate the probability that it belongs to class C. Using Bayes' classic theorem, we have:

$$p(C\mid F_1,...,F_n)=\frac{p(C)p(F_1,...,F_n\mid C)}{p(F_1,...,F_n)} \quad (1)$$

To simplify, we use the naïve Bayes classifier so that features $F_1,...,F_n$ are independent. Here we adapt the classic bag-of-words theory of Mooney and Roy [22] and regard a trace as an independent bag of actions, neglecting the logical relationship among the actions. Based on this assumption we have:

$$p(F_1,...,F_n\mid C)=p(F_1\mid C),...,P(F_n\mid C)$$
$$p(F_1,...,F_n)=p(F_1),...,p(F_n) \quad (2)$$

(1) is reformulated as:

$$p(C\mid F_1,...,F_n)=\frac{p(C)p(F_1\mid C),...,P(F_n\mid C)}{p(F_1),...,p(F_n)} \quad (3)$$

In our case, we aim at evaluating a user's competency on a certain concept with a trace he/she left on a set of concepts. So we adapt (1) as (4):

$$p(C_j\mid T_i)=\frac{p(C_j)p(T_i\mid C_j)}{p(T_i)} \quad (4)$$

where $p(C_j)$ is defined as the a priori probability that a random user has the highest competency on concept j of total J concepts. $p(C_j\mid T_i)$ represents the probability that a user i leaving trace T_i in the platform has the highest competency on concept j among all the N users. $p(T_i)$ is the probability that a user leaves a trace like T_i. As described previously, the trace of a user is a combination of actions on a variety of concepts. We define $p(T_i)$ in (5):

$$p(T_i)=p(A_{i,1})\times\cdots\times p(A_{i,n})=\prod_{k=1}^{n}p(A_{i,k}) \quad (5)$$

where $p(A_{i,k})$ represents the probability that actions of trace i on concept k happen. T_i is composed of actions on n concepts respectively. So (4) becomes:

$$p(C_j \mid T_i) = \frac{p(C_j)p(T_i \mid C_j)}{\prod_{k=1}^{n} p(A_{i,k})} \quad (6)$$

$p(C_j)$ is a constant because with no other constraints, all the users have the same probability to perform best at a certain concept. An estimate $\hat{p}(C_j)$ for $p(C_j)$ is:

$$\hat{p}(C_j) = \frac{1}{N} \quad (7)$$

where N is the total number of users. We measure users' performances by the frequency of actions. We define $p(A_{i,k})$ as the top percentage of rank of frequency among all users. Thus the higher the frequency of actions that user i takes on concept j ranks, the smaller $p(A_{i,k})$ is. For example, user No. 9 has applied actions which ranks 2 out of a group of 10 users, then $p(A_{9,j})$ = 0.2. It means if we randomly choose a user x from this set of users, the probability that x performed at least equally as user No.9 is 0.2.

$p(T_i \mid C_j)$ represents the probability of user i having a trace T_i if user i has the most competency on concept j. Two factors influence this value. Firstly, if a user has the most competency on j, it is highly probable that this user has much competency on concepts semantically nearby. As T_i is composed of a set of actions $\{A_{i,k} \mid A_{i,k} \in T_i\}$, we evaluate the semantic distance between k and j. We use $\omega_{k,j}$ to represent the weight of concept k on j. Fig. 4 shows a part of ontology of a use case for developing a semantic website. In view of complexity of calculations, we consider only the concepts semantically 2 edges away from j. Suppose j is the concept "Ontologic_request." Obviously, "Language" and "SQL" are two edges from j and we put their weight of influence to j as ω. "Request" and "SPARQL" are given 2ω and finally for the concept j itself we designate 4ω. The sum of weight of concepts is 10ω which is equal to 1. Secondly, given the weight between concept k and j, the higher user i ranks on concept k, the larger $p(T_i \mid C_j)$ is. We define:

$$p(T_i \mid C_j) = \frac{1}{Z} \sum_{\{k \mid A_{i,k} \in T_i\}} [1 - p(A_{i,k})] \times \omega_{k,j} \quad (8)$$

where Z is a scaling normalizing factor depending only on $\{A_{i,k} \mid A_{i,k} \in T_i\}$, that is, a constant if the values of the feature variables are known.

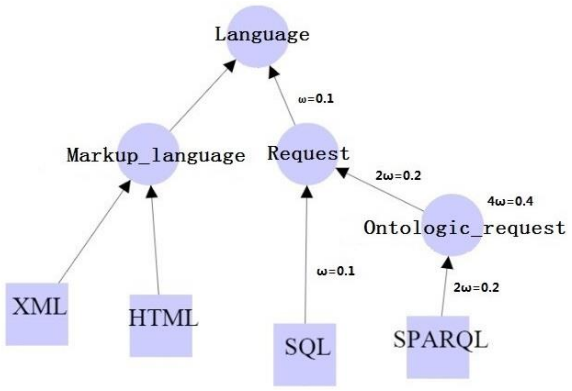

Figure 4. A part of ontology of a use case for developing a semantic website

We have:

$$p(C_j \mid T_i) = \frac{\sum_{\{k \mid A_{i,k} \in T_i\}} [1 - p(A_{i,k})] \times \omega_{k,j}}{N \times Z \times \prod_{k=1}^{n} p(A_{i,k})} \quad (9)$$

Finally, we obtain $p(C_j \mid T_i)$ and by comparing the probability of all users on the same concept, we can finally give a recommendation about who is most probably the "best" at a concept given his/her traces.

IV. EXPERIMENTATION

We apply our methods on the MEMORAe platform [23]. It allows users to collaborate by sharing different types of resources such as documents, notes, weblinks, etc. All these resources are organized in a semantic web of concepts they belong to. Fig. 5 shows the interface of MEMORAe.

The scenario taken is a five-user group collaborating on the theme "Star wars." We collect and analyze traces on the ontology of "Star wars" shown in Fig. 6. We illustrate the frequencies of all types of actions (addition, access, etc.) among all users on concerned concepts respectively by a histogram as shown in Fig. 7. For each concept, we demonstrate the proportion of each type of actions by each user by a pie chart as in Fig. 8. We define the weight of creation, access and deletion as 1, 0.8 and 0.2 as the importance declines.

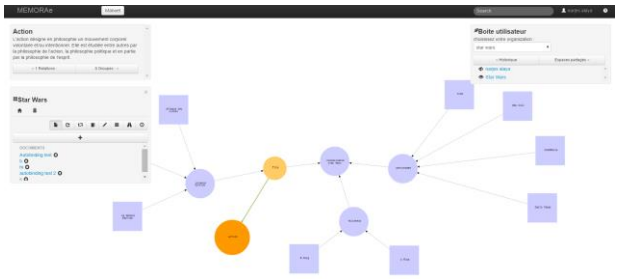

Figure 5. The interface of MEMORAe

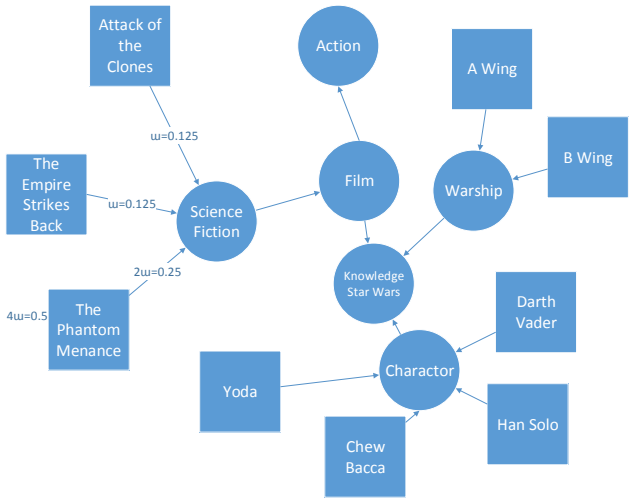

Figure 6. A part of the ontology of "Star Wars"

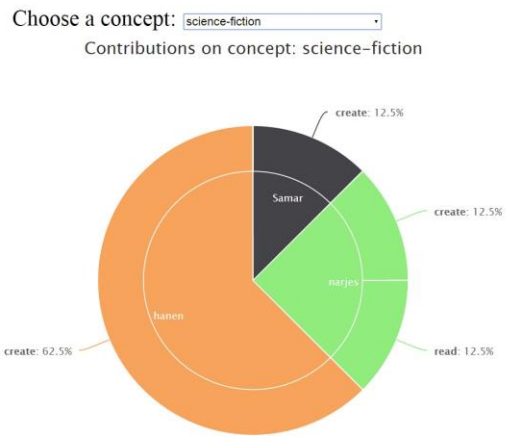

Figure 7. Pie chart of proportion of types of actions on "science fiction"

TABLE I. TRACES OF USERS ON CONCEPTS FROM ONTOLOGY OF USE CASE

	Frequency of Actions/Rank			
Name	**The Empire Strikes Back**	**Attack of the Clones**	**The Phantom Menance**	**Science fiction**
Marie Hélène	16/1	13/2	7/2	0/5
Samar	0/5	0/5	0/5	1/3
Narjes	6/3	6/3	3/3	2/2
Hanen	6/3	29/1	12/1	5/1
Cristian	0/5	0/5	0/5	0/0

To calculate the probability of "being most competent" on "The phantom menace" of "Hanen," we have:

$$p(C_{ThePhantomMenance} \mid T_1) = \frac{p(C_{ThePhantomMenance})\,p(T_1 \mid C_{ThePhantomMenance})}{\prod_{k=1}^{n} p(A_{1,k})} = \frac{31.25}{Z}$$

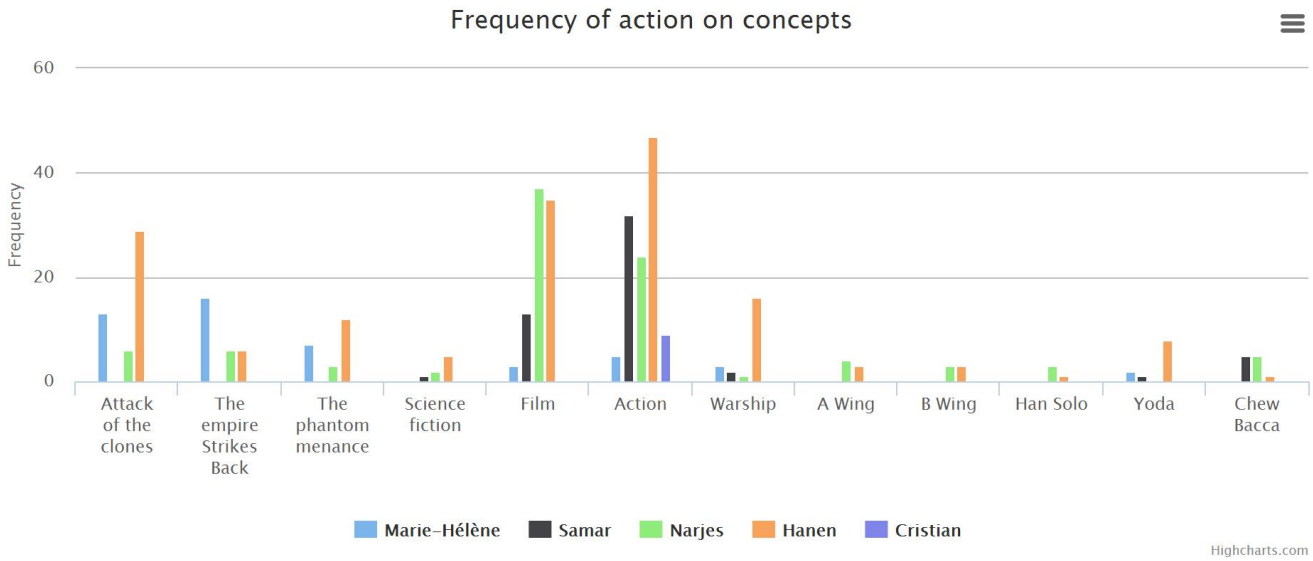

Figure 8. Frequencies of actions of users on varius concepts

TABLE II. TRACES OF USERS ON CONCEPTS FROM ONTOLOGY OF USE CASE

User Name	Probability for competency on "The Phantom Menance"
Marie Hélène	8.436%
Samar	0.1093%
Narjes	3.658%
Hanen	87.97.97%
Cristien	0

Results of calculations are shown in Table II in which "Hanen" earns an absolute high probability for being the most competent so that she is recommended as an expert on "The Phantom Menance." We observe that Hanen not only contributes the most for this concept, but also does a relatively good job in neighbouring concepts. It is the reason why she has an absolute high probability of competency on this concept, even if she is not the highest contributor on "The Phantom Menance" comparing to others. From this view, the approach values especially the importance of semantic relations between concepts.

V. CONCLUSION AND FUTURE WORK

A full exploitation of traces helps us organizing and improving collaboration. In this article we proposed a model of competency and a semantic structure to record traces. Secondly we proposed recommendations based on the evaluation of traces using Bayes Classifier. Finally, we orchestrated these methods to evaluate the competencies of users. We illustrated our method by an example. Results meet our expectation showing that this approach takes good care of semantic relations between concepts

Future work will include implementing our proposal of recommender system. Our previous work includes solving similar problem with TF-IDF. Testing methodology is needed to compare this method with our previous work and with other models.

REFERENCES

[1] J. Grudin, "Groupware and social dynamics: Eight challenges for developers." Communications of the ACM 37.1, pp. 92-105, 1994.

[2] D.M. Wegner, "Transactive memory: A contemporary analysis of the group mind." *Theories of group behavior*. Springer New York, pp. 185-208, 1987.

[3] J.R. Austin, "Transactive memory in organizational groups: the effects of content, consensus, specialization, and accuracy on group performance." *Journal of Applied Psychology* 88.5, pp. 866, 2003.

[4] R. Zarka, A. Cordier, E. Egyed-Zsigmond, and A. Mille, "Trace replay with change propagation impact in client/server applications." In IC 2011, 22èmes Journées francophones d'Ingénierie des Connaissances, pp. 607-622, 2012

[5] L. F. Tomaz, J. A.R. Nt, J.M. Souza, and G.B. Xexéo, "Bringing knowledge into recommendation systems." In *Computer Supported Cooperative Work in Design (CSCWD), 2011 15th International Conference on*, IEEE, pp. 246-252, 2011.

[6] N. Chang, M. Irvan, and T. Terano, "A tv program recommender framework." *Procedia Computer Science* 22, pp. 561-570, 2013.

[7] G. Adomavicius, and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." *Knowledge and Data Engineering, IEEE Transactions on* 17, no. 6, pp. 734-749, 2005

[8] Y.J. Chen, H.C. Chu, Y.M. Chen, and C.Y. Chao, "Adapting domain ontology for personalized knowledge search and recommendation." *Information & Management* 50, no. 6, pp. 285-303, 2013.

[9] S. Sahay, A. Venkatesh, and A. Ram, "Cobot: Real time multi user conversational search and recommendations." *Recommender Systems & the Social Web*, 2009.

[10] J.S. Breese, S. John, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering." In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 43-52. Morgan Kaufmann Publishers Inc., 1998.

[11] M.K. Condliff, D.D. Lewis, D. Madigan and C. Posse, "Bayesian mixed-effects models for recommender systems." In *ACM SIGIR*, vol. 99, pp. 23-30. 1999.

[12] D.M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles. "Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach." In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., pp. 473-480, 2000.

[13] N. Wang, M.H. Abel, J.P. Barthes, and E. Negre, "Towards a Recommender System from Semantic Traces for Decision Aid", *KMIS*, 2014.

[14] M. Ghazanfar, and P. Adam, "An Improved Switching Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering," 2010.

[15] A.I. Schein, A. Popescul, L.H. Ungar, and D.M. Pennock. "Methods and metrics for cold-start recommendations." *In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 253-260, 2002.

[16] P. Melville, R.J. Mooney, and R. Nagarajan. "Content-boosted collaborative filtering for improved recommendations." In *AAAI/IAAI*, pp. 187-192, 2002.

[17] Q. Li, M.H. Abel, and J.P. Barthès. "Facilitating Experience Sharing Groups-Collaborative Trace Reuse and Exploitation." In *International Conference on Knowledge Management and Information Sharing*, pp. 21-30, 2012.

[18] G. Antoniou, and F. Van Harmelen, *A semantic web primer*. MIT press, 2004.

[19] A. Molina, and M. Flores. "A virtual enterprise in Mexico: From concepts to practice." *Journal of intelligent and Robotic systems* 26, no. 3-4 , pp. 289-302, 1999.

[20] E. Müller, "Production planning and operation in competence-cell-based networks." *Production Planning & Control* 17, no. 2, pp. 99-112, 2006.

[21] X. Boucher, S. Peillon, and P. Burlat, "Towards a decision support for a collaborative increase of competencies within networks of firms." *Ecole de Mines de Saint-Etienne, Research Report G21-EMSE*, pp. 600-009, 2005

[22] R.J. Mooney, and L. Roy, "Content-based book recommending using learning for text categorization." In *Proceedings of the fifth ACM conference on Digital libraries*, ACM, pp. 195-204, 2000.

[23] M.H. Abel, B. Ahcene, L. Dominique, and M. Claude, "E-MEMORAe: a content-oriented environment for e-learning." In *E-Learning Networked Environments and Architectures*, Springer London, pp.186-205, 2007.