

EXAMEN Module UMINR 306 « Fouille de données »
3 novembre 2006

Tous documents autorisés

Durée : 2h

Chaque article est à TRAITER SUR UNE FEUILLE SEPARÉE

Vous devez choisir trois articles parmi les 4 qui sont proposés et répondre à l'ensemble des questions associées.

You have to address all the questions of three papers. Please use a separate sheet of paper for each article.

Article 1

"Clustering Data Streams: Theory and Practice" - Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani, Liadan O'Callaghan *IEEE Transactions on Knowledge and Data Engineering* - May/June 2003 (Vol. 15, No. 3) pp. 515-528

1. Résumez en quelques lignes l'article (notamment l'approche proposée en paragraphe 3).
Summarize the problem addressed in this paper and the main ideas of the proposed solution (paragraph 3).
2. Expliquez quelles sont les caractéristiques des data streams et pourquoi les méthodes classiques de clustering ne sont pas adaptées.
Explain what are the specificities of data streams and why classical clustering methods cannot be directly applied.
3. Y a-t-il des spécificités à la recherche de clusters par rapport aux approches liées aux fréquents présentées en cours ?
What are the specificities of clustering with respect to frequent pattern mining from data streams?
4. Que faut-il modifier pour construire des clusters flous ? Quelles conséquences cela aurait-il ?
How could this approach be extended to fuzzy clustering? What would be the consequences?

Article 2

"Mining the Web for synonyms: PMI-IR versus LSA on TOEFL" Turney, P.D *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, Freiburg, Germany, pp. 491-502.

1. Résumez en quelques lignes l'article de Peter Turney.
Summarize the problem addressed in this paper and the main ideas of the proposed solution.
2. L'algorithme PMI-IR consiste à déterminer la synonymie des mots. Cette méthode est-elle est généralisable pour la tâche de construction de classifications conceptuelles ?
PMI-IR is an algorithm that computes the synonyms of given words. Can this algorithm be applied to the task of conceptual classification?
3. Donnez les avantages et les limites de l'algorithme PMI-IR et de LSA. Quelles informations linguistiques peuvent être ajoutées aux algorithmes PMI-IR et LSA ?
Give the advantages and limits of the PMI-IR and LSA algorithms.
4. Décrivez brièvement la manière d'associer ces informations linguistiques aux différents algorithmes.
What kind of linguistic informations can be integrated to the PMI-IR and LSA algorithms? Give a brief description of how these linguistic informations can be integrated to the previous algorithms.

Article 3

"Efficient Mining of Temporally Annotated Sequences" Fosca Giannotti, Mirco Nanni and Dino Pedreschi. *Proceedings of the sixth SIAM International Conference on Data Mining (SDM'06)*, Bethesda, USA, 2006.

1. En quoi consiste les séquences annotées temporellement ?
What are temporally annotated sequences?
2. Quels sont les bénéfices de cette méthode ?
What are the benefits of this method?
3. Que proposeriez-vous pour aider l'utilisateur à analyser les résultats ?
What would you propose to help in the mining result analysis?
4. Les auteurs signalent dans leur conclusion qu'ils souhaitent étendre leurs travaux aux contextes spatiaux et spatiaux temporels. Donnez un exemple pertinent de base de données pouvant être traité par une telle approche, et des exemples de motifs extraits. Que doit-on modifier de cette approche pour trouver de tels motifs ?
The authors claim in their paper that it is possible to extend their approach to spatio and spatio-temporal data. Give an example of a database that could be addressed in this context and of the patterns that would be extracted. What is to be modified in this approach to deal with such spatio and spatio-temporal patterns?

Article 4

"Mining for Outliers in Sequential Databases" Pei Sun, Sanjay Chawla, Bavani Arunasalam (Best Paper) *Proceedings of the sixth SIAM International Conference on Data Mining (SDM'06)*, Bethesda, USA, 2006.

1. Résumez en quelques lignes l'article.
Summarize the problem addressed in this paper and the main ideas of the proposed solution.
2. Comment définiriez-vous la notion d'outlier ? Selon vous, cette notion est-elle liée à la notion d'exception ?
How would you define what an outlier is? Would you say that there is any relation with exceptions?
3. La recherche d'outliers s'appuie-t-elle nécessairement sur la notion de fréquents ?
Is the process of mining for outliers linked to the process of mining for frequent patterns ?
4. Les séquences présentées dans cet article sont-elles identiques à celles présentées en cours ? Quelles sont les différences ? Comment utiliser cette méthode dans le cadre des séquences étudiées en cours (décrivez les principales étapes à effectuer).
Are the sequences being considered in this paper the same as the ones presented in the course? Which are the differences? How would you use the method from this paper for mining the sequences that have been studied in the course? (describe the steps of the work to be done)