Mining Maximally Informative k-Itemsets in Massively Distributed Environments
Saber Salah, Reza Akbarinia, Florent Masseglia

To cite this version:
Saber Salah, Reza Akbarinia, Florent Masseglia. Mining Maximally Informative k-Itemsets in Massively Distributed Environments. BDA: Gestion de Données - Principes, Technologies et Applications, Nov 2016, Poitiers, France. lirmm-01411190

HAL Id: lirmm-01411190
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01411190
Submitted on 7 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives| 4.0 International License
Mining Maximally Informative k-Itemsets in Massively Distributed Environments

Saber Salah*
Inria and LIRMM, Zenith team,
University of Montpellier, France,
saber.salah@inria.fr

Reza Akbarinia
Inria and LIRMM, Zenith team,
University of Montpellier, France,
reza.akbarinia@inria.fr

Florent Masseglia
Inria and LIRMM, Zenith team,
University of Montpellier, France,
florent.masseglia@inria.fr

ABSTRACT
The discovery of informative itemsets is a fundamental building block in data analytics and information retrieval. While the problem has been widely studied, only few solutions scale. This is particularly the case when i) the data set is massive, calling for large-scale distribution, and/or ii) the length of the informative itemset to be discovered is high. In this paper, we address the problem of parallel mining of maximally informative k-itemsets (miki) based on joint entropy. We propose PHIKS (Parallel Highly Informative K-ItemSet) a highly scalable, parallel miki mining algorithm. PHIKS renders the mining process of large scale databases (up to terabytes of data) succinct and effective. Its mining process is made up of only two efficient parallel jobs. With PHIKS, we provide a set of significant optimizations for calculating the joint entropies of miki having different sizes, which drastically reduces the execution time of the mining process. PHIKS has been extensively evaluated using massive real-world data sets. Our experimental results confirm the effectiveness of our proposal by the significant scale-up obtained with high itemsets length and over very large databases.

La découverte d’itemsets informatifs est un élément fondamental dans l’analyse de données et la recherche d’information. Bien que le problème a été largement étudié, il y a peu de solutions qui passent à l’échelle. Ceci est particulièrement le cas lorsque i) les données sont de très grande taille, ce qui demande une distribution à grande échelle, et / ou ii) la longueur des itemsets informatifs à découvrir est élevée. Dans cet article, nous abordons le problème de la fouille des itemsets (miki) les plus informatifs (appelé miki) qui est calculé en considérant l’entropie conjointe des items. Nous proposons PHIKS (Parallel Highly Informative K-itemset), un algorithme parallèle d’extraction de miki. PHIKS rend le processus d’extraction de grandes bases de données à grande échelle (jusqu’à plusieurs téraoctets de données) rapide et efficace. Son processus d’extraction est constitué de seulement deux jobs parallèles. Avec PHIKS, nous proposons un ensemble d’optimisations significatives pour le calcul des entropies conjointes de miki ayant des tailles différentes, ce qui réduit considérablement le temps d’exécution du processus. PHIKS a été largement évalué en utilisant de grands jeux de données réels. Nos résultats expérimentaux confirment l’efficacité de notre proposition qui passe à l’échelle pour de grandes valeurs de k et sur de très grandes bases de données.

1. THE PHIKS APPROACH
Featureset, or itemset, mining [1] is one of the fundamental building bricks for exploring informative patterns in databases. Features might be, for instance, the words occurring in a document, the score given by a user to a movie on a social network, or the characteristics of plants (growth, genotyp, humidity, biomass, etc.) in a scientific study in agronomics. A large number of contributions in the literature has been proposed for itemset mining, exploring various measures according to the chosen relevance criteria. The most studied measure is probably the number of co-occurrences of a set of features, also known as frequent itemsets [2]. However, frequency does not give relevant results for a various range of applications, including information retrieval [3], since it does not give a complete overview of the hidden correlations between the itemsets in the database. This is particularly the case when the database is sparse [4]. Using other criteria to assess the informativeness of an itemset could result in discovering interesting new patterns that were not previously known. To this end, information theory [5] gives us strong supports for measuring the informativeness of itemsets. One of the most popular measures is the joint entropy of an itemset. An itemset X that has higher joint entropy brings up more information about the objects in the database.

We study the problem of Maximally Informative k-Itemsets (miki for short) discovery in massive data sets, where informativeness is expressed by means of joint entropy and k is the size of the itemset [6, 7, 8]. Miki are itemsets of interest that better explain the correlations and relationships in the data. Example 1 gives an illustration of miki and its potential for real world applications such as information retrieval.

EXAMPLE 1. In this application, we would like to retrieve documents from Table 1, in which the columns $d_1$, $d_{10}$ are documents, and the attributes A, B, C, D, E are some features (items, keywords) in the documents. The value “1” means that the feature occurs in the document, and “0” not. It is easy to observe that the itemset $(D, E)$ is frequent, because features $D$ and $E$ occur together in almost every document. However, it provides little help for document retrieval. In other words, given a document $d_1$ in our data set, one might look for the occurrence of the itemset $(D, E)$ and, depending on whether it occurs or not, she will not be able...
to decide which document it is. By contrast, the itemset \((A, B, C)\) is infrequent, as its member features rarely or never appear together in the data. And it is troublesome to summarize the value patterns of the itemset \((A, B, C)\). Providing it with the values \((1, 0, 0)\) we could find the corresponding document \(d_5\); similarly, given the values \((0, 1, 1)\) we will have the corresponding document \(d_3\). Although \((A, B, C)\) is infrequent, it contains lots of useful information which is hard to summarize. By looking at the values of each feature in the itemset \((A, B, C)\), it is much easier to decide exactly which document they belong to. \((A, B, C)\) is a maximally informative itemset of size \(k = 3\).

Miki mining is a key problem in data analytics with high potential impact on various tasks such as supervised learning [9], unsupervised learning [10] or information retrieval [3], to cite a few. A typical application is the discovery of discriminative sets of features, based on joint entropy, which allows distinguishing between different categories of objects. Unfortunately, it is very difficult to maintain good results, in terms of both response time and quality, when the number of objects becomes very large. Indeed, with massive amounts of data, computing the joint entropies of all itemsets in parallel is a very challenging task for many reasons. First, the data is no longer located in one computer, instead, it is distributed in parallel is a very challenging task for many reasons. First, the number of iterations of parallel jobs would be linear to over several machines. Second, the number of iterations of parallel data is no longer located in one computer, instead, it is distributed in parallel is a very challenging task for many reasons. First, the number of iterations of parallel jobs would be linear to over several machines. Second, the number of iterations of parallel data is no longer located in one computer, instead, it is distributed in parallel.

<table>
<thead>
<tr>
<th>Features</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(d_1)</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Features in the documents

PHIKS has been extensively evaluated using massive real-world data sets. Our experimental results show that PHIKS significantly outperforms alternative approaches, and confirm the effectiveness of our proposal over large databases containing for example one Terabyte of data.

2. ADDITIONAL AUTHORS

3. REFERENCES


