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 k 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 donnés 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 grane taille, ce qui demande une distribution à grande échelle, et / ou ii) la longueur k des itemsets informatifs à découvrir est élevée. Dans cet article, nous abordons le problème de la fouille des k iems 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 évaluée 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.

*Saber Salah - This work was partially supported by the Inria Project Lab Hemera.

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, genotype, 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_x 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*

Features	Documents									
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
A	1	1	1	1	1	0	0	0	0	0
B	0	1	0	0	1	1	0	1	0	1
C	1	0	0	1	0	1	1	0	1	0
D	1	0	1	1	1	1	1	1	1	1
E	1	1	1	1	1	1	1	1	1	1

Table 1: Features in the documents

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_3*; similarly, given the values* $(0, 1, 1)$ *we will have the corresponding document* d_6*. 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 over several machines. Second, the number of iterations of parallel jobs would be linear to k (i.e., the number of features in the itemset to be extracted [7]), which needs multiple database scans and in turn violates the parallel execution of the mining process. We believe that an efficient *miki* mining solution should scale up with the increase in the size of the itemsets, calling for cutting edge parallel algorithms and high performance evaluation of an itemset's joint entropy in massively distributed environments.

We propose a deep combination of both information theory and massive distribution by taking advantage of parallel programming frameworks such as MapReduce [11] or Spark [12]. To the best of our knowledge, there has been no prior work on parallel informative itemsets discovery based on joint entropy. We designed and developed an efficient parallel algorithm, namely Parallel Highly Informative K-itemSet (PHIKS in short), that renders the discovery of *miki* from a very large database (up to Terabytes of data) simple and effective. It performs the mining of *miki* in two parallel jobs. PHIKS cleverly exploits available data at each mapper to efficiently calculate the joint entropies of *miki* candidates. For more efficiency, we provide PHIKS with optimizations that allow for very significant improvements of the whole process of *miki* mining. The first technique estimates the upper bound of a given set of candidates and allows for a dramatic reduction of data communications, by filtering unpromising itemsets without having to perform any additional scan over the data. The second technique reduces significantly the number of scans over the input database of each mapper, i.e., only one scan per step, by incrementally computing the joint entropy of candidate features. This reduces drastically the work that should be done by the mappers, and thereby the total execution time.

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

[1] J. Han, *Data mining : concepts and techniques*. Elsevier/Morgan Kaufmann, 2012.

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 1994, pp. 487–499.

[3] E. Greengrass, "Information retrieval: A survey," 2000.

[4] H. Heikinheimo, E. Hinkkanen, H. Mannila, T. Mielikäinen, and J. K. Seppänen, "Finding low-entropy sets and trees from binary data," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007, pp. 350–359.

[5] T. M. Cover, *Elements of information theory*. Hoboken, N.J: Wiley-Interscience, 2006.

[6] R. Gray, *Entropy and information theory*. New York: Springer, 2011.

[7] A. J. Knobbe and E. K. Y. Ho, "Maximally informative k-itemsets and their efficient discovery," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006, pp. 237–244.

[8] C. Zhang and F. Masseglia, "Discovering highly informative feature sets from data streams," in *Database and Expert Systems Applications*, 2010, pp. 91–104.

[9] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proceedings of International Conference on Emerging Artificial Intelligence Applications in Computer Engineering*, 2007, pp. 3–24.

[10] Z. Ghahramani, "Unsupervised learning," in *Advanced Lectures on Machine Learning*, 2004, pp. 72–112.

[11] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[12] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proceedings of the 2Nd USENIX Conf. on Hot Topics in Cloud Computing*, 2010, pp. 10–10.