POLITECNICO DI TORINO
Repository ISTITUZIONALE

Discovering cross-topic collaborations among researchers by exploiting weighted association rules

Original
Discovering cross-topic collaborations among researchers by exploiting weighted association rules / Cagliero, Luca; Garza, Paolo; Kavoosifar, Mohammad Reza; Baralis, Elena. - In: SCIENTOMETRICS. - ISSN 0138-9130. - STAMPA. - 116:2(2018), pp. 1273-1301. [10.1007/s11192-018-2737-3]

Availability:
This version is available at: 11583/2711667 since: 2021-04-03T15:13:10Z

Publisher:
Springer

Published
DOI:10.1007/s11192-018-2737-3

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
Springer postprint/Author's Accepted Manuscript

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/s11192-018-2737-3

(Article begins on next page)

23 July 2024

Noname manuscript No.
(will be inserted by the editor)

Discovering cross-topic collaborations among researchers by exploiting weighted association rules

Luca Cagliero · Paolo Garza · Mohammad Reza Kavoosifar · Elena Baralis

Received: date / Accepted: date

Abstract Identifying the most relevant scientific publications on a given topic is a well-known research problem. The Author-Topic Model (ATM) is a generative model that represents the relationships between research topics and publication authors. It allows us to identify the most influential authors on a particular topic. However, since most research works are co-authored by many researchers the information provided by ATM can be complemented by the study of the most fruitful collaborations among multiple authors.

This paper addresses the discovery of research collaborations among multiple authors on single or multiple topics. Specifically, it exploits an exploratory data mining technique, i.e., Weighted Association Rule (WAR) mining, to analyze publication data and to discover correlations between ATM topics and combinations of authors. The mined rules characterize groups of researchers with fairly high scientific productivity by indicating (i) the research topics covered by their most cited publications and the relevance of their scientific production separately for each topic, (ii) the nature of the collaboration (topic-specific or cross-topic), (iii) the name of the external authors who have (occasionally) collaborated with the group either on a specific topic or on multiple topics, and (iv) the underlying correlations between the addressed topics.

The applicability of the proposed approach was validated on real data acquired from the Online Mendelian Inheritance in Man (OMIM) catalog of genetic disorders and from the PubMed digital library. The results confirm the effectiveness of the proposed strategy.

Keywords Author Topic Model · Weighted Association Rule Mining · Data Mining and Knowledge Discovery

Politecnico di Torino - Dipartimento di Automatica e Informatica - Torino, Italy
E-mail: {luca.cagliero,paolo.garza,mohammadreza.kavoosifar,elena.baralis}@polito.it

1 Introduction

Nowadays, most scientific publications such as conference proceedings, scientific journal, and books are accessible through digital libraries and online databases. For example, in genetics and genomics PubMed (NCBI, 2017) and OMIM (Hamosh et al, 2000) are among the most popular publication repositories. Researchers commonly perform manual topic- or author-driven queries on publication data to retrieve the content of interest. However, this activity can be extremely time consuming and prone to errors, because the number of publications to explore may be large.

Automatically discovering the most relevant publications related to a given topic is a well-known data mining problem, which has already been addressed in literature (Rosen-Zvi et al, 2012). For example, the Author-Topic Model (ATM) is an established generative model which can be exploited to represent authors' interests. It first analyzes the textual content of publication documents to characterize latent topics as probability distributions over words. Then, topics are associated with most influential authors. In most related works (e.g., Ding et al (2014); Kim et al (2016); Rosen-Zvi et al (2012); Tang et al (2008)) the reputation of an author in the research community is derived from the popularity of his publication. For example, an established way to measure the relevance of a publication in the research community is to count the number of received citations (Lu et al, 2015). A thorough overview of the related literature is given in Section 2.

Most research works are the result of collaborations among multiple authors. Teams of researchers typically produce a large body of publications related to specific topics. Furthermore, researchers may collaborate with external research teams on complementary topics. However, the ATM is, to the best of our knowledge, unable to identify fruiful research collaborations among multiple authors. Furthermore, the underlying correlations between multiple topics are unknown. Solving these issues is particularly challenging because it requires correlating the contributions of multiple authors on multiple topics by evaluating the significance of their joint research studies with respect to the existing literature. Therefore, there is a need for automated data mining solutions aimed to analyze publication data and to identify fruitful research collaborations among multiple authors.

This paper addresses the problem of discovering cross-topic collaborations among multiple authors by means of an exploratory data mining technique, i.e., weighted association rule mining (Wang et al, 2000). Specifically, it analyzes publication data and topics to discover interesting patterns, called Weighted Association Rules (WARs). WARs represent recurrent implications between combinations of authors and/or topics. Topics can be either described by publication metadata (if available) or automatically inferred by ATM. WARs characterize the activities of groups of authors (of arbitrary size) that have produced a set of relevant publications. They are extracted only if they hold for many highly cited publications. For each group of researchers WARs answer to the following questions:

(1) On what topics is the collaboration among researchers focused on?
(2) Is the collaboration focused on a specific topic or spread over multiple topics?
(3) What is the relevance of their scientific production separately for each topic?
(4) Has the group (occasionally) collaborated with external authors? On which topics?
(5) To what extent are the topics addressed in the collaboration correlated with each other?

Depending on their characteristics, WARs may answer to one or more of the questions above. Furthermore, WARs can be easily ranked by decreasing relevance to simplify the exploration of the mining result. As discussed in Section 2, this work is, to the best of our knowledge, the first attempt to exploit WARs (Wang et al, 2000) to analyze cross-topic collaborations among authors.

We experimentally evaluated the effectiveness of the proposed approach on data acquired from two independent libraries, i.e., OMIM (Hamosh et al, 2000) and PubMed (NCBI, 2017), which collect genomic and genetic studies. Specifically, we analyzed publications available in OMIM enriched with citation counts crawled from PubMed. To discover fruitful collaborations among researchers working together on specific genetic disorders we extracted WARs by considering as topics the metadata descriptions associated with each OMIM publication. In parallel, we automatically extracted also a description of the ATM main topics from each publication document and then we discovered WARs representing correlations between authors and ATM topics. The results show that the mined WARs allow experts to gain insights into the analyzed data. Specifically, WARs of different categories allow experts to effectively face complementary issues and to answer to different research questions. Furthermore, the quality indices associated with WARs allow us to rank the discover patterns based on their relative significance thus easing the manual exploration of the mining result.

The rest of the paper is organized as follows. Section 2 compares the proposed approach with existing studies. Section 3 thoroughly describes the proposed methodology, while Section 4 experimentally evaluates its effectiveness on real data. Finally, Section 5 draws conclusions and discusses future developments of the proposed work.

2 Related work

This work is partly related to the following research topics: (i) Author-Topic Model, (ii) Graph-based co-authorship models, (iii) Citation content analysis, (iv) Reviewer assignment, and (v) Weighted association rule mining. Hereafter, we will separately overview each topic and discuss the position of our work with respect to existing studies.

Author-Topic Model. The problem of modeling the interests of authors on different topics based on textual document analysis has already been investigated in literature. The Author-Topic Model (ATM) (Rosen-Zvi et al, 2012) is a generative model for textual documents, where topics are represented as probability distributions over words while authors are associated with probability distributions over topics. The ATM allows us to represent the original documents as a mixture of topics and to determine which authors have mainly contributed to a given topic. For example, given a set of publications the corresponding research topics can be extracted first. Then, the subset of most active researchers on each topic can be extracted. Steyvers et al (2004) have proposed a Bayesian approach to estimate the ATM parameters. Since the ATM correlates single authors with specific topics, it cannot be directly applied to infer cross-topic collaborations among multiple authors.

Graph-based co-authorship models. Graph- and network-based models have already been adopted to model co-authorship relationships (e.g., Mutschke (2003); Newman (2001); Tang et al (2008); White and Smyth (2003)). Specifically, in (Mutschke, 2003; Newman, 2001) graph theory and visualization models have jointly been exploited to model co-authorship and citation relations. White and Smyth (2003) used a graph indexing technique (i.e., PageRank (Brin and Page, 1998)) to identify the most authoritative researchers. The relationships among researchers can be also modeled as social networks. For example, ArnetMiner (Tang et al, 2008) is a social network of academic researchers, where for each author a research profile is automatically extracted from the Web and integrated with publication data accessible through existing digital libraries. Network- and graph- models represent connections between authors without explicitly considering the correlations with the covered topics. Therefore, the underlying information differs from those provided by the patterns considered in this study.

Citation content analysis. To study the impact of scientists' research, the number of citations received by their scientific publications has been considered in several studies (e.g., Ding et al (2014); Kim et al (2016); Zhang et al (2013)). Citation content analysis is the research branch that focuses on studying citations among papers thus computing a reputation score for each researcher. Specifically, it focuses on analyzing the semantics, syntax, and position in the text of the paper of the citations to reveal the influence of both authors and scientific papers. For example, Kim et al (2016) analyzed the sentences including citation expressions to identify interesting characteristics of scholarly communication. Ding et al (2014) and Zhang et al (2013) classified citations based on their semantics to gain insights into the relationships between authors and topics. In our work, citations are exploited to weigh the relevance of a publication thus, indirectly, to measure the reputation of a group of researchers related to a given topic. However, our analysis is not focused on citation analysis. As discussed in Section 3, to measure the relevance of a publication different measures can be easily integrated as well.

Reviewer assignment. A related branch of research concerns the assignments of reviewers to scientific papers. The aim is to support editors in the

peer review of scientific papers by automatically recommending potential reviewers. For example, Kou et al (2015a,b); Li and Hou (2016) addressed the problem of choosing a pool of reviewers for a given paper based on the expertise of a potentially large set of candidate reviewers and on the main topics covered by the paper under review. The authors tackled the optimization problem to assign each paper to at least three independent reviewers with complementary expertise so that the pool of reviewers assigned to each paper covers most of the topics addressed by the paper and each reviewer has a reasonable number of reviews to do. Unlike the works proposed by Kou et al (2015a,b); Li and Hou (2016) the task addressed in this work is not an optimization problem. The techniques adopted in our work are exploratory and allow us to discover interpretable patterns useful for supporting a number of advanced analyses.

Weighted association rule mining. A parallel research effort has been devoted to efficiently extracting itemsets and association rules from weighted data (Cagliero and Garza, 2014; Sun and Bai, 2008; Tao et al, 2003; Wang et al, 2000). This problem extends the traditional association rule mining task, which was first introduced by Agrawal et al (1993) in the context of market basket analysis, to the case in which data items are no longer considered as equally relevant within the analyzed data. For example, in the context of market basket analysis the goal is to find sets of products frequently purchased together by taking into account not only the list of products that customers have put into their market basket but also the purchased amount and unitary price of each purchased product. Wang et al (2000) proposed to extract weighted association rules, i.e., rule including weights denoting item significance are extracted. Tao et al (2003) and Cagliero and Garza (2014) used weights to drive the frequent and infrequent itemset mining processes, respectively, while Sun and Bai (2008) automatically generated weights by means of graph indexing techniques. This work focuses on extracting weighted association rules from publication data to discover cross-topic collaborations among authors. A preliminary version of this work has been presented by Cagliero et al (2017). This work extends its preliminary version to a large extent. The main differences can be summarized as follows:

(i) Topics are characterized as probability distributions over words which are automatically extracted from publication documents and not only selected from publication metadata.

(ii) Weighted Association Rules (WARs), which represent implications between combinations of authors and topics, are extracted as well on top of frequent itemsets. The newly extracted patterns measure the strength of an implication between authors and topics (e.g., to what extent the citations received by a group of researchers are related to a specific topic) and not only the observed frequency of appearance of a combination of authors in the publication dataset. To the best of our knowledge, this work is the first attempt to exploit WARs to analyze cross-topic collaborations among authors.

(iii) WARs are classified based on their goal into five main categories. WAR categories allow us to identify not only topic-specific collaborations but also *cross-topic* collaborations among authors.

3 Cross-topic Scientific collaboration analyzer

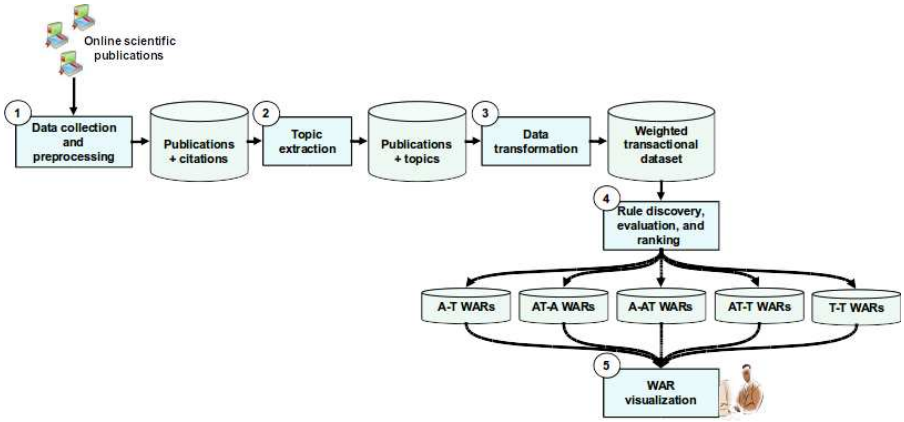

Fig. 1 Cross-topic Scientific Collaboration Analyzer

Cross-topic Scientific Collaboration Analyzer (CSCA) is a data-driven methodology to automatically discover significant cross-topic collaborations among authors of scientific publications. The methodology relies on the application of an exploratory data mining technique to publication data acquired from digital libraries or online databases such as PubMed (NCBI, 2017) and OMIM (Hamosh et al, 2000).

The aim is to identify groups of authors who have significantly contributed to the research community related to a particular topic or to a given set of topics. The relevance of the scientific production of a group is evaluated by considering the number of citations received by the co-authored publications. For each identified group CSCA extracts, classifies, and ranks patterns, called Weighted Association Rules (WARs), that allow us to answer to the following questions:

(1) On what topics is the collaboration among researchers focused on?

(2) Is the collaboration focused on a specific topic or spread over multiple topics?

(3) What is the relevance of their scientific production separately for each topic?

(4) Has the group (occasionally) collaborated with external authors? On which topics?

(5) To what extent are the topics addressed in the collaboration correlated with each other?

The methodology consists of five main steps, which are depicted in Figure 1:

(i) *Data collection and preprocessing.* Publications data and related metadata are acquired from online sources, preprocessed to make them suitable for the

next mining process, and then stored into a centralized repository (see Section 3.1).
(ii) *Topic extraction.* The topics covered by each publication are extracted from either publication metadata or from the textual content of the publication by exploiting the Author-Topic Model (see Section 3.2).
(iii) *Data transformation.* Author information, citation counts, and publication topics are prepared to the association rule mining step (see Section 3.3).
(iv) *Rule discovery, evaluation, and ranking.* Weighted Association Rules (WARs), which represent implications between combinations of authors and topics, are extracted, classified, and ranked to support knowledge discovery from publication data (see Section 3.4).
(v) *Rule visualization.* The mined WARs are visualized through a Web-based application to ease result exploration. The interface allows experts to constrain both the type and the content of the visualized WARs (see Section 3.5).

A more thorough description of each step follows.

3.1 Data collection and preprocessing

Publication data are acquired from digital libraries and online databases (e.g., PubMed (NCBI, 2017), OMIM (Hamosh et al, 2000)) by exploiting the exposed Application Programming Interfaces (APIs) and then stored in a unique repository.

For our purposes, for each publication we acquire the following data:
(i) the Digital Object Identifier (DOI) of the publication,
(ii) the list of authors,
(iii) the current number of citations received,
(iv) the text of the publication, and
(v) any relevant (domain-specific) metadata associated with the publication.

The current number of citations is considered as one of the main indicators of influence/popularity of a scientific publication in the research community (Lu et al, 2015). Hereafter, we will consider it as reference indicator of the influence/popularity of a publication. However, since the proposed methodology is general, different measures can be easily integrated as well (e.g., the Hirsch index (Hirsch, 2010)). A more thorough discussion on the choice of the most appropriate citation counting method is given by (Waltman and van Eck, 2015).

Publication data can be enriched with metadata describing the addressed topics. For example, the OMIM database (Hamosh et al, 2000) collects publications about genomics and genetics. For each publication the list of related genes and genetic disorders is given. As discussed in Section 3.2, we will consider such information (if available) to identify the main topics covered by the publication.

To prepare publication data to the next mining processes, the text of the publications and the related metadata are cleaned by applying two established text preprocessing steps:

Stopword elimination. Stopword elimination filters out the words having least semantic content, because their presence would bias the quality of the next mining phase. Specifically, the words occurring in the text are compared with those contained in a dictionary of conjunctions, articles, prepositions, abbreviations etc. To focus the next topic extraction phase on the most significant document content matching words are removed. To perform stopword elimination, in our experiments we used the Natural Language Toolkit (NLTK) stopword corpus (Loper and Bird, 2002).

Stemming. Stemming is an established text preprocessing technique whose aim is to reduce words to their root form (i.e., the stem) (Tan et al, 2005). This step reduces the variance of the textual content to a more compact set of word roots. For example, nouns, verbs in gerund form, and past tenses (e.g. words *correlation*, *correlated*, *correlating*) are remapped to a common root form (e.g. *correlat*).

Furthermore, the author names and the descriptors of genes and genetic disorders are made uniform by removing noisy characters, abbreviated forms, etc.

3.2 Topic extraction

To each publication a list of covered topics is assigned. Depending on the data source, topics can be either described by metadata (e.g., genes and genetic disorders in the OMIM database (Hamosh et al, 2000)) or unknown. We propose two complementary strategies to assign topics to each publication: if topic metadata are given, CSCA exploits metadata content as descriptors of the covered topic. Furthermore, CSCA will extract a description of the main topic covered by each publication document by exploiting the Author-Topic Model (ATM) (Rosen-Zvi et al, 2012).

ATM is a generative model for textual documents, where documents in the input collection are mixture of topics. Each topic is a probability distribution over word stems as described in the Latent Dirichlet Allocation (LDA) model (Blei et al, 2003). More specifically, for each publication document a distribution over topics is first sampled from a Dirichlet distribution. Next, for each word stem in the document a single topic is assigned according to the distribution. Finally, each word stem is sampled from a multinomial distribution over word stems specific to the sampled topic (Rosen-Zvi et al, 2012). In the computation, the generative algorithm keeps track of a $W \times T$ (word stem-by-topic) and a $A \times T$ (author-by-topic) count matrices. The algorithm starts by assigning word stems to random topics and authors from the set of authors and documents. Count matrices are stored from 10 samples (with random initial assignments) at the 2000th iteration of the Gibbs sampler. From the count matrices topics and authors are extracted. Each topic is characterized by (i) word-based description W_{de}, i.e., the top-10 word stems that are most likely to be generated conditioned on the topic, and (ii) author-based

description A_{de}, i.e., the top 10 most likely authors to have generated a word stem conditioned on the topic.

For each publication document we extract the top-k main topics by following the procedure described by Algorithm 1. We scan the input document to find the word stems that are included in the description W_{de} of any topic in T. For each topic we store the maximum per-word count in $W \times T$ over all words in its description. Since word counts indicate the relevance of word in the topic, we assign the top-k topics associated with the word stems with maximal count.

Algorithm 1 Main topic detection

Require: the publication documents D, the word stem-by-topic count matrix $W \times T$, and the word stem descriptions W_{de} of all topics T
Ensure: set of main topics $t^* \in T$ for each document d in D
for all d in D **do**
  $top[t]=0 \ \forall \ t \in T$
  for all word stem w occurring in D **do**
    for all topic t in T **do**
      if $w \in W_{de}$ **then**
        update $top[t]$ if the w's count in $W \times T$ is higher than the current $top[t]$ value
      end if
    end for
  end for
  select the top-k topics of d associated with the k maximal values in top
end for
return the top-K topics of each document d

3.3 Data transformation

Publication data, citation scores, and topics are stored into a weighted transactional dataset. A weighted transactional dataset is a set of pairs ⟨transaction, weight⟩, where each *transaction* corresponds to a different scientific publication, while *weight* is the value of the citation counter for the corresponding publication (see Section 3.1).

Transactions consist of sets of items, where items are publication authors (e.g., *Smith, L.*), or research topics (e.g., *topic X*). Topics can be described either by metadata content or by ATM description (see Section 3.2). Items are represented in the form (*feature:value*), where *feature* is *Author* or *Topic*, while *value* is the corresponding feature value.

A more formal definition of weighted transactional dataset is given below.

Definition 1 Weighted transactional dataset. Let A be the set of authors and T be the set of topics. Let P be the set of all scientific publications and let $C(p_i)$ ($p_i \in P$) be an influence/popularity score associated with publication p_i. An item i_k is a pair *feature:*v_q, where $v_q \in A$ if *feature* is equal to *Author* or $v_q \in T$ if *feature* is equal to *Topic*. A transaction t_j is a set of items related to publication p_j. A weighted transactional dataset $\mathcal{D}$ is a set of

Table 1 Example of weighted transactional dataset

Pub. id	Citation count	Authors	Topics
1	10	(*Author*:*Brown, J.*), (*Author*:*Smith, L.*)	(*Topic*:*A*), (*Topic*:*X*), (*Topic*:*Z*)
2	5	(*Author*:*Brown, J.*), (*Author*:*Smith, L.*)	(*Topic*:*D*), (*Topic*:*X*)
3	10	(*Author*:*Brown, J.*), (*Author*:*Smith, L.*)	(*Topic*:*C*), (*Topic*:*Z*)
4	1	(*Author*:*Smith, L.*)	(*Topic*:*X*), (*Topic* : *Z*)
5	10	(*Author*:*Brown, J.*), (*Author*:*Smith, L.*)	(*Topic*:*C*) (*Topic*:*X*)
6	12	(*Author*:*Smith, L.*)	(*Topic*:*Z*)

weighted transactions, where each weighted transaction $tw_j \in \mathcal{D}$ corresponds to a different publication $p_j \in P$ and it consists of a pair $\langle t_j, C(p_j)\rangle$.

For instance, Table 1 reports an example of dataset consisting of six weighted transactions, each one corresponding to a different scientific publication. Each publication, identified by the respective id, is weighted by the corresponding number of citations (see Column *Citation count*). For each publication the list of authors (see Column *Authors*) and the covered topics (see Column *Topics*) are known. Publications can be co-authored, and can be related to many topics. For example, publication with pub. id 1 received 10 citations (i.e., transaction weight equal to 10). Its corresponding transaction consists of the following items: *Author*:*Brown, J.*, *Author*:*Smith, L.*, *Topic*:*A* and *Topic*:*X*. The transaction refers to a publication that was co-authored by Brown J. and Smith L. and that relates to topics A and X.

3.4 Pattern discovery, evaluation, and ranking

This step entails applying an exploratory data mining approach, i..e, Weighted Association Rule (WAR) mining, to the prepared weighted transactional dataset. The aim is to automatically generate patterns, i.e., the WARs, representing interesting implications between combinations of authors and topics. WARs are then classified based on their semantic meaning into three main categories and ranked to simplify the manual exploration of the mining result.

This section is organized as follow. Section 3.4.1 introduces the concept of WAR and its quality indices, Section 3.4.2 provides a high-level description of the algorithm used to extract the WARs of interest. Finally, Section 3.4.2 introduces the WAR categories and discusses how they can be exploited to help experts to answer to the research questions introduced in Section 1.

3.4.1 Weighted association rules

Association rule mining (Agrawal et al, 1993) is an established data mining technique to discover recurrent correlations among data items hidden in large datasets. Association rule mining is commonly performed as a two-step process which entails (i) frequent itemset mining from the transactional data and

(ii) association rule discovery from the set of frequent itemsets mined at the previous step.

Frequent itemset mining. A k-itemset is a set of k distinct items in a transactional dataset. It indicates the co-occurrence of the corresponding items in the analyzed dataset. In our context of analysis, an item represents either an author or a topic (see Definition 1). Hence, itemsets may represent co-occurrences of multiple authors and topics in the analyzed dataset. A more formal definition of itemset is given below.

Definition 2 Itemset. Let $\mathcal{D}$ be a weighted transactional dataset and let $\mathcal{I}$ be the set of distinct items in the form *feature*:v_q contained in any weighted transaction $tw_j \in \mathcal{D}$. A k-itemset (i.e., an itemset of length k) is a set of k distinct items in $\mathcal{I}$.

Note that each itemset may contain an arbitrary number of items belonging to any feature.

Since generating all the possible itemsets is computationally intractable even on medium-size datasets, itemset mining is commonly driven by a minimum support threshold (Agrawal et al, 1993). More specifically, frequent itemset mining entails extracting all the itemsets that *frequently* occur in the source dataset $\mathcal{D}$, i.e., all itemsets whose frequency of occurrence (support) in the source dataset is above a given threshold *minsup*. The support threshold prevents the extraction of less relevant or misleading itemsets. Thus, it allows us to consider only the most recurrent and thus potentially reliable patterns.

For example, itemset $\{(Author{:}Brown,\ J.),(Topic{:}X)\}$ occurs three times in the dataset in Table 1 (publications with ids 1, 2, and 5). Hence, by enforcing a minimum support threshold *minsup*=2 the itemset would be extracted because its frequency of occurrence (3) is above the minimum (user-provided) threshold.

Unfortunately, the number of frequent itemsets can be very large. To prevent the generation of redundant patterns, thus simplifying the manual inspection of the result, a more compact subset of frequent itemsets, called the closed itemsets (Wang et al, 2003), can be extracted. An itemset is *closed* if there exists no superset that has the same support as this original itemset.

Itemset evaluation based on weighted support. The support quality index of an itemset does not consider the relative importance of each transaction in the source dataset (Agrawal et al, 1993). More specifically, in our context of analysis, each publication may have a different impact on the research community. Some publication can be highly influential, whereas others may have a limited scope. Hence, to evaluate pattern significance pattern occurrences in each publication are weighted according to its impact on the research community.

Since our goal is to generate only the combinations of authors and topics that have achieved a high impact, we extended the standard itemset mining problem by integrating item weights (Wang et al, 2000). Specifically, item occurrences within each transaction (publication) are weighted by a influence/popularity score, such as the citation count (see Section 3.1). Therefore,

the co-authorship of publications with a large number of citations is rewarded, whereas co-authorship of publications with few citations are penalized. To formalize this step, we introduce the concept of *weighted support* of an itemset as a weighted frequency of occurrence of the itemset in the weighted transactional dataset.

Definition 3 Weighted support of an itemset. Let $\mathcal{D}$ be a weighted transactional dataset and I be an itemset. Let tw_j: $\langle t_j,\ C(p_j)\rangle$ be an arbitrary weighted transaction in $\mathcal{D}$. The weighted support of I in $\mathcal{D}$, hereafter denoted by wsup(I), is defined as follows:

$$wsup(I) = \sum_{tw_j \in \mathcal{D} | I \subseteq t_j} C(p_j)$$

Recalling the previous example, $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.),(Topic{:}X)\}$ has a weighted support equal to 25 because it covers the weighted transactions with publication ids 1 (weight 10), 2 (weight 5), and 5 (weight 10), respectively.

Weighted association rule discovery. Weighted Association Rules (WARs) are extracted on top of frequent itemsets. Given two itemsets A and B (of arbitrary length) a weighted association rule $A \rightarrow B$ is an implication between A and B. A more formal definition follows.

Definition 4 Weighted association rule. Let A and B be two itemsets. A weighted association rule is represented in the form $R : A \rightarrow B$, where A and B are the body and the head of the rule respectively.

A and B are also denoted as antecedent and consequent of rule $A \rightarrow B$. Association rule extraction is commonly driven by weighted support (wsup) and confidence (wconf) quality indexes (Agrawal et al, 1993). While the weighted support index represents the weighted frequency of occurrence of the rule in the source dataset, the weighted confidence index represents the rule strength.

Definition 5 Weighted support of a WAR. Let $\mathcal{D}$ be a weighted transactional dataset. The weighted support (wsup) of a weighted association rule $R : A \rightarrow B$ is defined as the weighted support of $A \cup B$ in $\mathcal{D}$.

Definition 6 Weighted confidence of a WAR. Let $\mathcal{D}$ be a weighted transactional dataset. The weighted confidence (wconf) of a weighted association rule $R : A \rightarrow B$ is the conditional probability of (weighted) occurrence in $\mathcal{D}$ of itemset B given itemset A, i..e,

$$wconf(R) = \frac{wsup(R)}{wsup(A)} = \frac{wsup(A \cup B)}{wsup(A)}$$

.

For example, WAR $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.)\} \rightarrow (Topic : X)\}$ indicates an implication between a couple of authors and a specific topic. The WAR has weighted support equal to 25 and weighted confidence equal to $\frac{25}{35}$, because the implication holds for publications with ids 1, 2, and 5 but not for publication with id 3 (citation count = 10).

WAR categories. For our purposes, we consider five main categories of WARs tailored to our context of analysis. Each category consists of the set of all WARs characterized by a predefined sequence of items (authors and/or topics). Categories are tailored to different research questions.

Category 1: Authors-Topic Rules. These rules are extracted because they allow us to answer to the following questions: *On what topics is the collaboration focused on? Is the collaboration focused on a specific topic or spread over multiple topics?*

WARs of category Authors-Topic (hereafter denoted as A-T WARs for the sake of brevity) are represented in the form $R : A \rightarrow B$, where the rule antecedent A is an arbitrary itemset consisting of a set of authors, while the consequent B is an arbitrary itemset including a single topic.

For example, $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.)\} \rightarrow (Topic : X)\}$ is an A-T WAR. It indicates that authors J. Brown and L. Smith have co-authored publications related to topic X. $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.)\} \rightarrow (Topic : Z)\}$ is another A-T WAR with the same antecedent, which indicates that the same authors have collaborated on topic Z. If both WARs are extracted, then the two authors have fruitfully collaborated on multiple topics in separate publications. Notice that if the same publications cover multiple topics, part of co-authored publications may cover both topics. We will separately consider this particular case in the WAR category 4 (see *AuthorsTopics-Topic Rules*).

The weighted support of the A-T WARs indicate the sum of the citation counts of all the co-authored publications on the given topic. Sorting rules by decreasing wsup allow experts to consider first the research collaborations that have received a fairly high attention from the research community. Notice that WARs with low wsup are early pruned during the mining process (due to support threshold enforcement), because the corresponding collaborations were very unlikely to produce significant results.

The weighted confidence indicates the fraction of citations received by the co-authored publications on the considered topic with respect to the total number of citations received by all the co-authored publications (independently of the topic). Sorting rules by decreasing wconf allows experts to select, among all the topics covered during the collaboration, the topics that have achieved the highest impact. A-T WARs with high wconf indicate the topics on which the collaboration is mainly focused on.

Given a combination of authors, the wsup index allows experts to filter out the less relevant collaborations. On the other hand, the wconf value indicates the strength of the correlation between the set of authors and a particular topic. For example, if the wconf of a A-T WAR is close to 100% (all the

citations are associated with a particular topic) then it means that the collaborations was productive only on the corresponding topic.

Category 2: AuthorsTopic-Author Rules. These rules are extracted because they allow us to answer to the following question: *Working on a given set of topics, has the group (occasionally) collaborated with external authors?*

WARs of category AuthorsTopic-Author (hereafter denoted as AT-A WARs for the sake of brevity) are represented in the form $R : A \rightarrow B$, where the rule antecedent A is an arbitrary itemset consisting of a set of authors and a set of topics, while the consequent B is an arbitrary itemset including a single author.

For example, WAR $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.), (Topic{:}X)\}$ $\rightarrow (Author{:}Black,\ J.)$ is an AT-A WAR. It indicates that in the collaboration between authors J. Brown and L. Smith on topic X they have collaborated with author J. Black. WAR $\{(Author{:}Brown,\ J.), (Topic{:}X)\ (Topic{:}Z)\} \rightarrow$ $(Author{:}Smith,\ L.)$ is another AT-A WAR which indicates a cross-topic collaboration between a couple of authors.

The weighted support of the AT-A WAR indicates the significance of the collaboration between the group under analysis and the external author. The weighted confidence indicates the impact of this collaboration on the productivity of the group of authors associated with the given topic. For example, if the wconf is 50% it means that half of the citations received by the combination of authors on the considered topic was achieved by works co-authored by the considered author. Therefore, low wconf value indicate occasional (yet potentially fruitful) collaborations, whereas high wconf values indicate more systematic collaborations between the group and external authors.

Category 3: Authors-AuthorTopic Rules. These rules are extracted because they allow us to answer to the following question: *Has the group collaborated with external authors? On which topics?*

WARs of category Authors-AuthorTopic (hereafter denoted as A-AT WARs for the sake of brevity) are represented in the form $R : A \rightarrow B$, where the rule antecedent A is an arbitrary itemset consisting of a set of authors, while the consequent B is an arbitrary itemset including a single author and a single topic.

For example, $\{(Author{:}Brown,\ J.),(Author{:}Smith,\ L.)\} \rightarrow \{(Author{:}Black,\ J.),$ $(Topic : X)\}$ is an A-AT WAR. It indicates that in the research works made in the collaboration between authors J. Brown and L. Smith the authors have frequently collaborated with author J. Black on topic X.

The weighted support of the AT-A WAR indicates the significance of the collaboration between the group of authors and the consider pair author-topic. The weighted confidence indicates the impact of this topic-specific collaboration on the overall productivity of the group of authors in the antecedent of the rule (independently of the topic). For example, if the wconf is 50% it means that half of the citations received by the combination of authors (independently of the topic) was achieved by works co-authored by the external author on the indicated topic. Low wconf values may be due either to the low

productivity of the collaboration between the group and the external authors or to the low popularity of the topic.

Category 4: AuthorsTopics-Topic Rules. These rules are extracted because they allow us to answer to the following question: *Given a group of researchers who have frequently collaborated on a set of topics, which other topic is likely to be covered by their co-authored publications?*

WARs of category AuthorsTopics-Topic (hereafter denoted as AT-T WARs) describe cross-collaborations between authors. Since in a collaboration each member could provide its expertise on a particular topic, it is interesting to investigate on which topics an existing author-topic collaboration could be specialized.

For example, $\{(Author{:}Brown,\ J.), (Author{:}Smith,\ L.), (Topic : X)\} \rightarrow \{(Topic : Z)\}$ is an AT-T WAR. It indicates that an authors' collaboration on topic X is frequently associated with an additional topic (Z).

If the wconf of the AT-T WAR is very high (close to 100%) most of the co-authored publications related to topic X cover topic Z as well. Hence, these rules allow us to measure the strength of the cross-topic authors' collaborations.

Category 5: Topics-Topic Rules. These rules are extracted because they allow us to answer to the following question: *To which topic is a particular set of topics most correlated with?* Since authors' collaborations are often cross-topic, analyzing the underlying correlation between multiple topics is particularly interesting.

For example, an example of Topics-Topic WARs (hereafter denoted as T-T WARs) is $\{(Topic : A), (Topic : X)\} \rightarrow \{(Topic : Z)\}$.

Sorting T-T WARs by decreasing confidence allows us to identify the sets of most correlated sets of topics.

3.4.2 The extraction algorithm

Many frequent Weighted Association Rule (WAR) mining algorithms have already been proposed in literature (e.g., Cagliero and Garza (2014); Sun and Bai (2008); Tao et al (2003); Wang et al (2000)). To accomplish the WAR mining task from weighted transactional data, we applied to a two-step mining process which entails (i) Closed itemset mining, and (ii) WAR generation from closed itemsets. Step (i) is accomplished by an FP-Growth-based algorithm (Han et al, 2000). The algorithm relies on an FP-tree data model, i.e., a compact, tree-based representation of the original dataset residing in main memory. Itemset extraction is optimized to generate only closed itemsets. Step (ii) focuses on generating WARs from closed itemsets by generating any combinations of closed itemsets representing WARs of interest (Agrawal and Srikant, 1994). WARs for all categories and topics are extracted in a single run by visiting in the FP-tree structure in a depth-first manner.

3.5 WAR visualization

WARs of different categories are visualized to ease expert validation. A Web interface allows domain experts to browse the rules belonging to the category of interest, to filter WARs not including any specific combinations of authors or topics, and to sort the extracted WARs by decreasing weighted confidence.

Thanks to the graphical interface, domain experts can more easily identify which author-topic combinations are potentially of interest for advanced analysis.

Figure 2 shows a screenshot of the developed interface. The interface can be accessed at following link: http://dbdmg.polito.it/CSCA/.

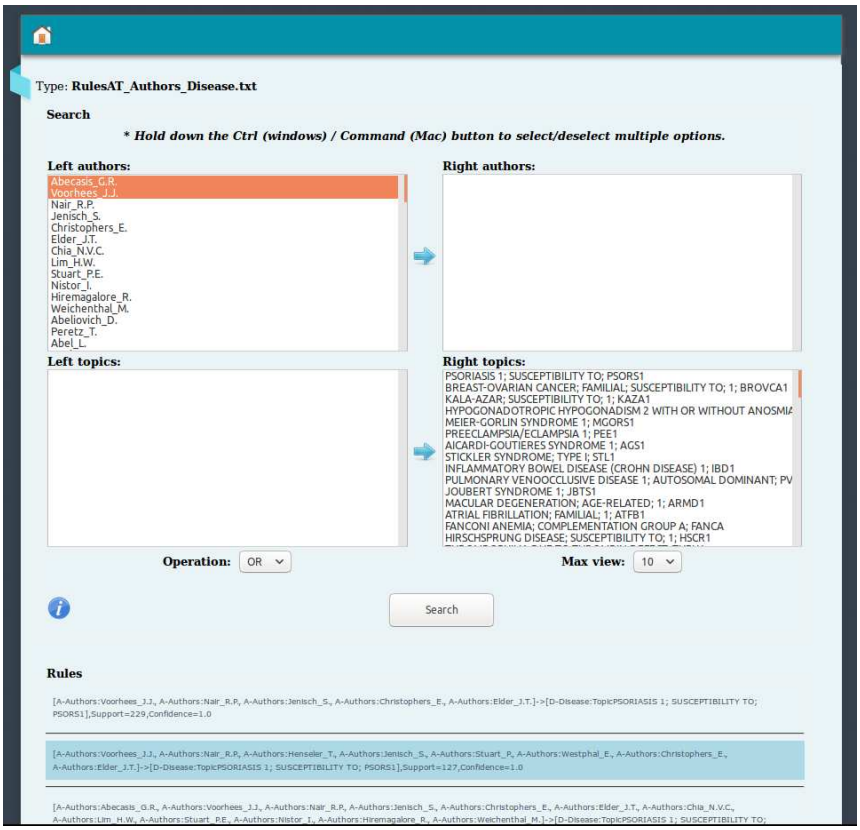

Fig. 2 The WAR visualization interface.

4 Experimental results

We studied the applicability of the proposed methodology in a real case study, i.e., the analysis of the research collaboration on genomics or genetics. To perform our experiments we analyzed publication data and citations acquired from the Online Mendelian Inheritance in Man (OMIM) catalog of genetic disorders (Hamosh et al, 2000). The aim is to discover from OMIM publication data and the related citations/topics the groups of researchers that have conducted the most influential studies on genomics or genetics.

The experiments were performed on a 2.30 GHz Intel Core workstation with 16 GB of RAM, running Ubuntu Linux 16.04 LTS. The data crawling, preprocessing, and preparation steps are based on Java programs, while the weighted association rule mining algorithm is written in C.

Data sources. The Online Mendelian Inheritance in Man (OMIM) database (Hamosh et al, 2000) is one of the most comprehensive and authoritative compendium of human genes and genetic phenotypes. OMIM is part of the National Center for Biotechnology Information (NCBI) system of databases (NCBI, 2017) and it is freely available on the Web. OMIM collects information on all known Mendelian disorders and over 12,000 genes. Specifically, it thoroughly describes the relationships between phenotypes and genotypes by providing full-text, referenced overviews on genetic disorders. The database is updated daily and thus its content is continuously evolving over time. OMIM exposes public Application Programming Interfaces (APIs) for genetic data crawling and download. Specifically, it allows users to acquire the list of all known disorders and a set of related annotations. Disorder annotations consist of (i) a list of scientific publications ranging over the disorder (for each publication the complete bibliographic information is known), (ii) a textual description of the disorder including references, and (iii) links to other genetics resources. To crawl data from the online OMIM database, we exploited the exposed APIs (Hamosh et al, 2000). To retrieve the number of citations received by each publication in OMIM we exploited the APIs of the PubMed digital library (NCBI, 2017). The integrated dataset, which were obtained by integrating publication data crawled from OMIM and citation data crawled from PubMed, contains 8825 articles, 34555 authors, and 302 disorders.

Prepared datasets. For each publication in OMIM topics can be extracted either from metadata (i.e., the descriptions of the genetic disorders associated with the publication) or from the Author-Topic Model (see Section 3.2). However, part of the OMIM publications have no full-text access through the exposed APIs. Therefore, we enriched all publications in OMIM with topics extracted from metadata, while we applied the ATM to extract 10 topics to the subset of the publications in OMIM for which the full-text is available. For the sake of brevity, hereafter we will denote as *Disorder* the dataset collecting OMIM publication, disorder topics, and citation counts, while we will denote as *ATM* the dataset collecting the portion of OMIM publication with free full-text version, the related citations, and the ATM main topics. For each paper of *Disorder*, one single disorder topic per paper is available.

Differently, for the *ATM* dataset, we selected the top 5 most related topics for each paper, based on the output of the ATM algorithm.

Comparison betwee OMIM disorders and ATM topics. We analyzed the similarity between the ten automatically extracted ATM topics and the manually assigned 302 OMIM disorders in the analyzed publication data. Specifically, we first analyzed the distribution of the OMIM disorders within each subset of publications related to the same topic. Most topics appeared to be almost uncorrelated with the OMIM disorders, as the most frequent disorders typically occurred in no more than 5% of the publications of a given topic. Furthermore, OMIM disorders are associated with 80%-90% of the ATM topics. Hence, the two categorizations seem to be not correlated with each other, as they were generated in different ways and with completely different purposes.

This section is organized as follows. Section 4.1 reports some examples of WARs belonging to different categories, which allowed us to answer to the research questions posed in the previous sections. In the subsquent sections, a quantitative analysis of the mining results is reported. Specifically, we discuss (i) the accuracy of the mined rules in identifying the main topics covered by a set of researchers (Section 4.2.1), (ii) the distribution of the extracted WARs in the selected categories (Section 4.2.2), (iii) the impact of the parameter settings on the number of extracted WARs (Sections 4.2.3-4.2.4) and (iv) the algorithm complexity and execution time (Section 4.2.5).

4.1 Knowledge discovery from the mined WARs

This section reports some examples of WARs separately for each category and shows us how these patterns can be exploited to answer to the questions posed in the previous sections (see Section 3).

Category (1) comprises Authors-Topic weighted rules (A-T WARs). They can be used to answer to the following questions:

On what topics each collaboration focused on?
Which are the most fruitful authors' collaborations?
Is the authors' collaboration focused on a specific topic or spread over multiple topics?

Table 2 reports the top 5 Authors-Topic rules (A-T WARs), in order of decreasing wsup, mined from Disorder. Each A-T rule indicates a specific set of authors who have profitably collaborated on a particular topic. Rule profitability was measured in terms of number of citations received by the co-authored publications. In fact, a high wsup value implies a high number of citations for the papers co-authored by the set of authors reported in the antecedent of the A-T rule. Based on the extracted WARs, we discover, for instance, that authors Siddique T. and Deng H. X. wrote a set of papers on the Amyotrophic lateral sclerosis disorder and their co-authored publications have

been cited 1861 times. Since this WAR is the most frequent one among all the mined A-T WARs ranging over the topic, we can deduce that Siddique T. and Deng H. X. are among the most influential/authoritative group of researchers about Amyotrophic lateral sclerosis.

Table 2 *Disorder dataset*: Top 5 Authors-Topic rules (A-T WARs) in terms of wsup

A-T rule	**wsup**	**wconf**
{(*Author*:Bignell, G.R.), (*Author*:Davies, H.), (*Author*:Garnett, M.J.), (*Author*:Cox, C.), (*Author*:Stephens, P.), (*Author*:Edkins, S.), (*Author*:Clegg, S.), (*Author*:Teague, J.), (*Author*:Woffendin, H.), (*Author*:Bottomley, W.), (*Author*:Davis, N.), (*Author*:Dicks, E.)} → {(*Topic*:MELANOMA CUTANEOUS MALIGNANT SUSCEPTIBILITY TO 1)}	1861	100%
(*Author*:Siddique, T.), (*Author*:Deng, H.-X.) → (*Topic*:AMYOTROPHIC LATERAL SCLEROSIS 1)	1828	100%
(*Author*:Hentati, A.), (*Author*:Siddique, T.), (*Author*:Deng, H.-X.) → (*Topic*:AMYOTROPHIC LATERAL SCLEROSIS 1)	1800	100%
(*Author*:Rioux, J.D.), (*Author*:Silverberg, M.S.) → (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)	1470	100%
(*Author*:Silverberg, M.S.), (*Author*:Barmada, M.M.) → (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)	1388	100%

As readers can notice from Table 2, the most frequent A-T WAR is associated with a relatively large group of authors, which consists of 12 different authors. This is typical in the medical domain for which papers are usually co-authored by a large number of authors.

All the WARs reported in Table 2 are characterized by maximal confidence value (100%). This means that the set of authors appearing in the rule antecedent have collaborated only on the topic reported in the consequent of the associated rule. For instance, Siddique T. and Deng H. X. have had fruitful collaborations on the Amyotrophic lateral sclerosis disease but they have not produced significant literature on any other topics (according to our data-driven analyses). However, the authors who have had fruitful collaborations on a specific topic are likely to collaborate on other topics as well. To investigate whether authors' collaborations are focused on a specific topic or spread over multiple topics we can compare the A-T WARs characterized by the same antecedent by considering their confidence values as well. For example, Table 3 reports four WARs that can be exploited to characterize the collaborations between Brown, E.M. and Kifor, O. and those between Seidman, J.G. and Seidman, C.. Specifically, the first two A-T WARs reported in Table 3 show that Brown, E.M. and Kifor, O. have had fruitful collaborations on two main topics: HYPOCALCIURIC HYPERCALCEMIA FAMILIAL TYPE I and HYPOCALCEMIA AUTOSOMAL DOMINANT 1. Their papers on the first topic have received 79.4% of their overall citations (by considering only the co-authored publications), while the second topic is associated with 20.6% of their citations. The sum of the two rule confidence values is 100%. Hence, Brown, E.M. and Kifor, O. have collaborated only on the

aforesaid topics. The last two rules reported in Table 3 can be used to characterize the collaborations between other two researchers. Based on the mined rules Seidman, J.G. and Seidman, C. have had profitable collaborations on two topics (ARDIOMYOPATHY FAMILIAL HYPERTROPHIC 1 and CARDIOMYOPATHY DILATED 1A). However, since the sum of the confidence values of those rules is less than 100%, we can deduce that Seidman, J.G. and Seidman, C. have co-authored papers on other topics as well, but the latter works have not received a sufficiently high number of citations to be deemed as "relevant" (i.e., no other A-T WARs with wsup above 50 and wconf above 50% associated with Seidman, J.G. and Seidman, C. were mined).

Table 3 *Disorder dataset*: Examples of A-T WARs describing authors who have collaborated on multiple topics

A-T rule	**wsup**	**wconf**
{(*Author*:Brown, E.M.), (*Author*:Kifor, O.)} → {(*Topic*:HYPOCALCIURIC HYPERCALCEMIA FAMILIAL TYPE I)}	485	79.4%
{(*Author*:Brown, E.M.), (*Author*:Kifor, O.)} → {(*Topic*:HYPOCALCEMIA AUTOSOMAL DOMINANT 1)}	126	20.6%
{(*Author*:Seidman, J.G.), (*Author*:Seidman, C.)} → {(*Topic*:CARDIOMYOPATHY FAMILIAL HYPERTROPHIC 1)}	566	52.8%
{(*Author*:Seidman, J.G.), (*Author*:Seidman, C.)} → {(*Topic*:CARDIOMYOPATHY DILATED 1A)}	196	18.3%

Table 4 *Disorder dataset*: Examples of AT-A WARs

AT-A rule	**wsup**	**wconf**
{(*Author*:Rioux, J.D.), (*Author*:Silverberg, M.S.), (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)} → {(*Author*:Barmada, M.M.)}	1385	94.2%
{(*Author*:Rioux, J.D.), (*Author*:Silverberg, M.S.), (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)} → {(*Author*:Bitton, A.)}	852	57.9%

Beyond identifying fruitful collaborations among researchers and the corresponding topics, experts can be interested in analyzing (occasional) collaborations between the aforesaid groups and other researchers. Specifically, we want to answer to the following question:

"*Working on a given topic, has the group (occasionally) collaborated with external authors?*".

The AuthorsTopic-Author rules (AT-A WARs) can support experts in tackling this issue. Table 4 reports two example AT-A WARs that can be used to discover who have collaborated with Rioux, J.D. and Silverberg, M.S. on topic INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1 disease (the interest of the group on the specific topic were previously discovered

by analyzing the fourth A-T WAR reported in Table2). According to the AT-A WAR, Rioux, J.D. and Silverberg, M.S. have conducted joint works with Barmada, M.M. and Bitton, A. on the INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1 topic. Specifically, 94.2% of their citations on that topic are due to papers co-authored by Barmada, M.M. as well, while Bitton, A. has co-authored papers associated with 57.9% of their citations.

Table 5 *Disorder dataset*: Examples of A-AT WARs

A-AT rule	**wsup**	**wconf**
{(*Author*:Almer, S.), (*Author*:Finkel, Y.)} → {(*Author*:Colombel, J.-F.), (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)}	67	6.2%
{(*Author*:Cho, J.H.), (*Author*:Brant, S.R.)} → {(*Author*:Bayless, T.M.), (*Topic*:INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1)}	140	15.4%

Experts could be interested in analyzing the collaborations between a group of researchers and "external" researchers and discovering the topics of these collaborations. Specifically we are interested in answering to the question:

"*Has the group (occasionally) collaborated with external authors? On which topics?*"

Table 5 reports some examples of Authors-AuthorTopic rules (A-AT WARs). They can be exploited to answer to these questions. Based on the mined rules, the group of authors Almer, S. and Finkel, Y. has frequently collaborated only with Colombel, J.-F. on the INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1 topic. Moreover, this collaboration has covered only the 6.2% of their total citations (independently of the topics of the papers co-authored by Almer, S. and Finkel, Y.). Hence, authors Almer, S. and Finkel, Y. seem to have had a limited collaborations with researches external to their group. The second rule reported in Table 5 shows the "external" collaborations of the set of authors Cho, J.H. and Brant, S.R.. Even this group of authors has had an 'external" collaboration with another researcher (Bayless, T.M.) and the target of the collaboration was the INFLAMMATORY BOWEL DISEASE - CROHN DISEASE - 1 topic.

When papers are characterized by multiple topics, experts can also analyze the AuthorsTopics-Topic rules (AT-T WARs) to characterize the cross-topic collaborations among authors. Specifically, we are interested in answering to the question:

"*Given a set of co-authors collaborating on a set of topics, which other topic is likely to be covered by their co-authored publications?*"

Since the Disorder dataset contains a single topic per paper, in the following we will consider the AT-T WARs extracted from the ATM dataset as representative example (see Table 7). For instance, based on the first rule re-

Table 6 *ATM dataset*: ATM topis

Topic ID	Top-10 most related terms
T0	rat, neuron, muscl, effect, dai, studi, calcium, group, activ, induc
T1	gene, mutat, express, sequenc, protein, develop, analysi, dna, cell, genet, genom
T2	respons, drug, increas, potenti, channel, membran, effect, studi, function, reduc
T3	cancer, associ, studi, breast, increas, case, model, genotyp, risk, smoke
T4	health, data, base, method, studi, model, system, develop, predict, approach
T5	brain, imag, memori, tissu, inject, studi, model, control, test, network
T6	infect, hiv, viru, associ, immun, vaccin, diseas, antigen, reactiv, hepat
T7	cell, express, activ, induc, tumor, human, regul, protein, mice, receptor
T8	protein, activ, cell, bind, fig, membran, acid, level, α, dna
T9	patient, studi, group, ag, risk, conclus, year, method, treatment, associ

Table 7 *ATM dataset*: Examples of AT-T WARs

AT-T rule	wsup	wconf
{(*Author*:Shelbourne, P.), (*Author*:Davies, J.), (*Author*:Johnson, K.), (*Topic*:T8)} $\rightarrow$ {(*Topic*:T9)}	466	100%
{(*Author*:Johnson, K.), (*Author*:Buxton, J.), (*Topic*:T6)} $\rightarrow$ {(*Topic*:T8)}	456	100%

ported in Table 7, we can state that 100% of the publications related to topic T8 co-authored by Shelbourne, P., Davies, J., Johnson, K., Shelbourne, P., Davies, J., and Johnson, K. cover also Topic T9. Hence, the publications of the reported co-authors related to topic T8 are also related to topic T9 (i.e., those publications are related to the cross-topic collaboration on topics T8 and T9). Similar considerations hold for the second example AT-T WAR. For the sake of completeness, Table 6 reports the top 10 most related terms extracted by the ATM algorithm for each of the identified topics.

Table 8 *ATM dataset*: Examples of T-T WARs

T-T rule	wsup	wconf
{(*Topic*:T3), (*Topic*:T5), (*Topic*:T6), (*Topic*:T8)} $\rightarrow$ {(*Topic*:T7)}	1326	95.5%
{(*Topic*:T2), (*Topic*:T5), (*Topic*:T8), (*Topic*:T9)} $\rightarrow$ {(*Topic*:T4)}	1449	93.4%
{(*Topic*:T2)} $\rightarrow$ {(*Topic*:T0)}	2118	27.8%
{(*Topic*:T5)} $\rightarrow$ {(*Topic*:T0)}	2205	23.8%

Independently of the authors, we could be interested in analyzing the correlations among multiple topics to understand if the same topics are frequently covered by the same publication. Specifically, we are interested in answering to the question:

"*Given a set of publications related to a particular subset of topics, which other topic is also frequently covered in those publications?*"

Table 8 reports some examples of Topics-Topic rules (T-T WARs), which can be used to identify frequent correlations among topics. Specifically, Table 8 reports the top two most confident T-T WARs mined from the ATM dataset and the two less confident ones. The mined WARs show that single topics are usually not very correlated with each other (i.e., the last two rules are both characterized a low confidence value), while the publications covering a large set of topics can be highly correlated with a further topic (see the first two rules reported in Table 8). This result is consistent with the main goal of the ATM algorithm, which aims at identifying orthogonal topics (i.e., couples of topics are likely to be weakly correlated). Based on the last two T-T WARs reported in Table 8, it turns out that T2 is not very correlated with T0, and T5 is almost uncorrelated with T0 as well. The aforesaid considerations are consistent with the results of a qualitative comparison between the corresponding word-based topic descriptions in Table 6.

4.2 Quantitative evaluation of the WAR mining process

The goal of this section is manifold. Specifically,
(i) We report a quantitative assessment of the reliability of the mined WARs on publication data (see Section 4.2.1).
(ii) We analyze the per-length and per-category WAR distributions by setting a standard configuration for the WAR mining algorithm (see Section 4.2.2).
(iii) We discuss the impact of the algorithm parameter settings on the quality of the mining results (see Section 4.2.3 and 4.2.4).
(iv) We discuss the complexity of the CSCA system and we evaluate system performance in terms of execution time (see Section 4.2.5).

4.2.1 Quantitative assessment of the correctness of the mined WARs

We performed a quantitative assessment of the reliability of the mined WARs. As a case study, we focused this validation phase on A-T WARs and, separately for each dataset, we selected the top 50 WARs by decreasing weighted confidence. The aim of this validation process is to estimate to what extent each of the mined rules is relevant by measuring the pertinence of the topic recommended by the rule head with those of the most influential studies of the authors indicated in the rule body. Specifically, for each A-T WAR r we compared the topic in the rule r's consequent with those of the top 3 most cited publications of each author in the rule antecedent. Then, we defined as *score* of rule r the percentage of authors who published at least one of his top cited publications on the rule topic. This measure indicates the extent to which the authors mentioned in the rule have the assigned topic in their expertise. A high rule score indicates that the co-occurrence between multiple authors and the topic, which were extracted from publication data based on citation counts, is unlikely to be generated by chance as they reflect the expected single author-topic dependencies.

The average score was 99.8% for the Disorder dataset and 97.5% for the ATM dataset, respectively. This result confirms that the extracted author-topic associations can be deemed as reliable.

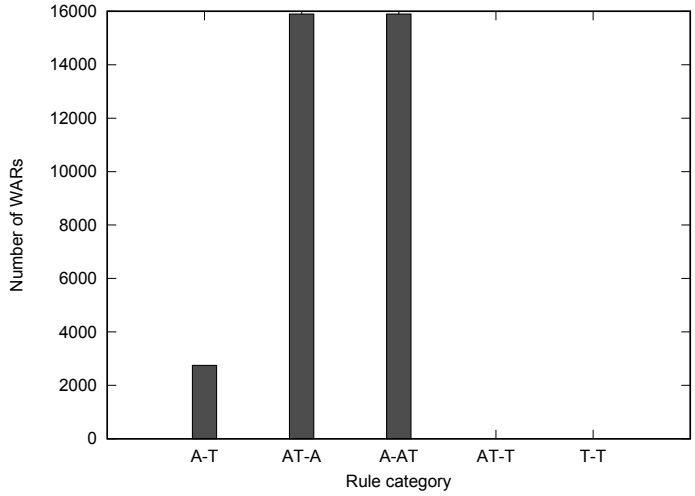

Fig. 3 *Disorder dataset*: Distribution of WARs per category. wsup=50, wconf=50%.

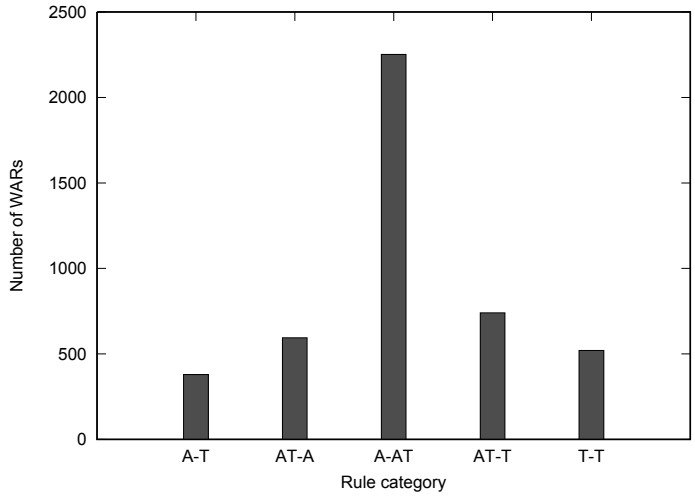

Fig. 4 *ATM dataset*: Distribution of WARs per category. wsup=50, wconf=50%.

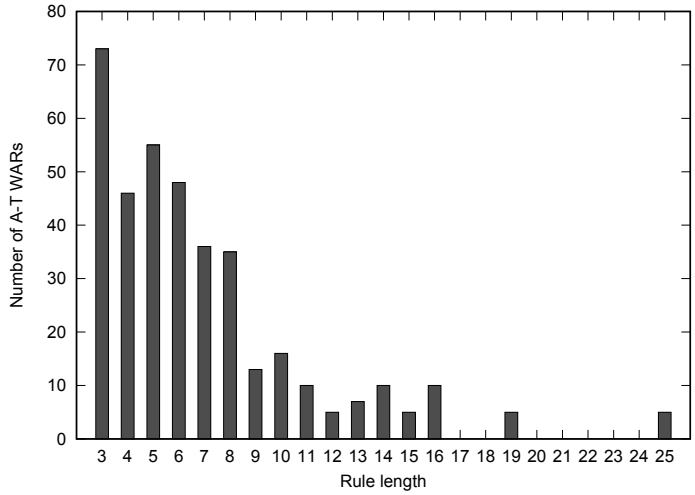

Fig. 5 *ATM dataset*: Distribution of A-T WARs per length. wsup=50, wconf=50%.

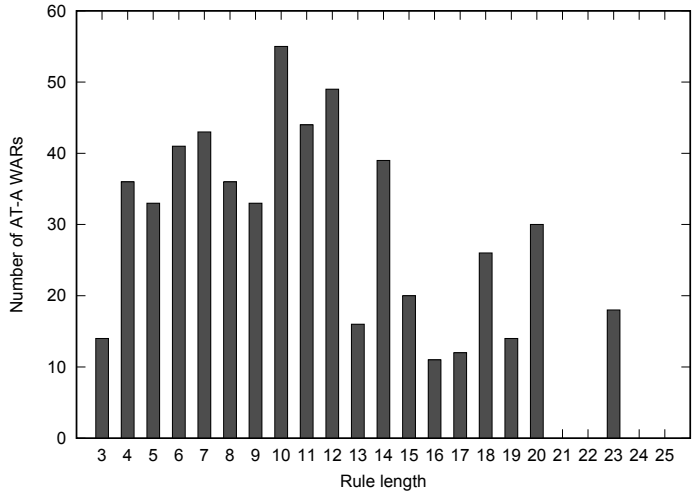

Fig. 6 *ATM dataset*: Distribution of AT-A WARs per length. wsup=50, wconf=50%.

4.2.2 Characteristics of the mined WARs

To analyze the characteristics of the mined WARs we first set, as standard configuration, the minimum weighted support threshold (i.e., the least citation count value) to 50 and the minimum weighted confidence threshold (i.e., the minimum percentage of publications for which the implication holds) to 50%. The impact of the aforesaid parameters will be discussed later.

Figures 3 and 4 respectively plot the number of WARs per category (see Section 3.4.1) mined from the *Disorder* and *ATM* datasets. As expected, the

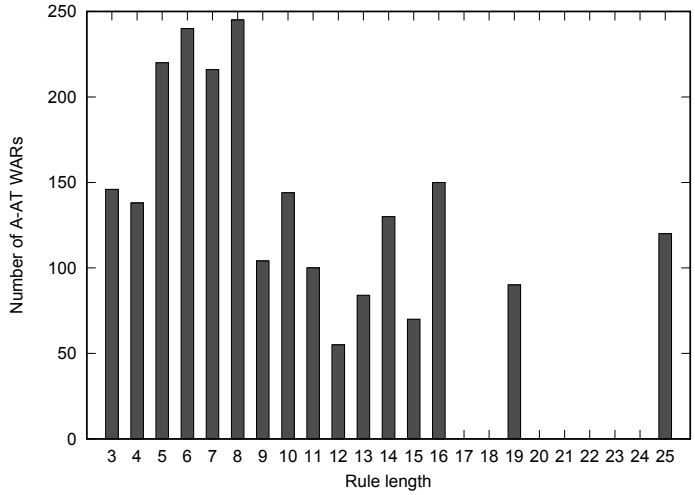

Fig. 7 *ATM dataset*: Distribution of A-AT WARs per length. wsup=50, wconf=50%.

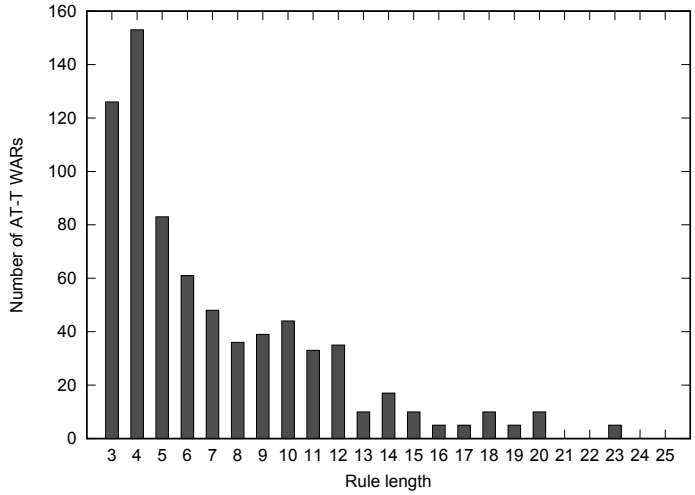

Fig. 8 *ATM dataset*: Distribution of AT-T WARs per length. wsup=50, wconf=50%.

number of A-T WARs is significantly lower than those of the other ones, because the number of possible combinations is usually at least one order of magnitude lower. The distributions of AT-A and A-TA WARs are approximately the same when only one topic per article is available (i.e., for the *Disorder* dataset) because they are generated from the same closed itemset by permuting the corresponding items.

For each category, we analyzed also the per-length distribution of the corresponding WARs (i.e., the number of contained items). As representative examples, Figures 5-9 report the per-length distribution of WARs of differ-

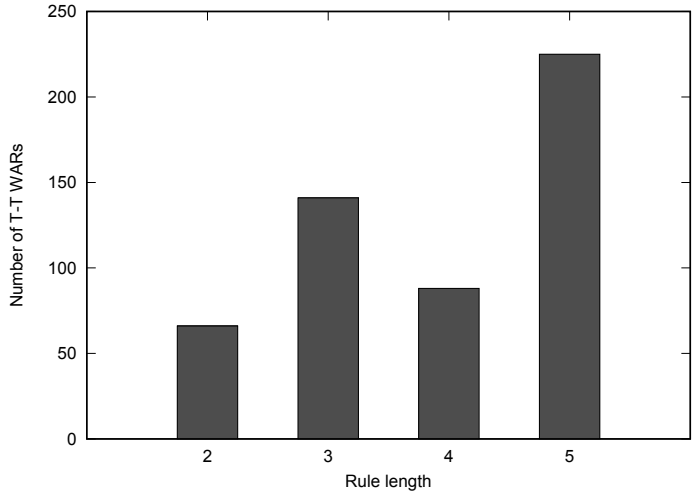

Fig. 9 *ATM dataset*: Distribution of T-T WARs per length. wsup=50, wconf=50%.

ent categories mined from *ATM*. We selected the rules mined from the *ATM* dataset because ATM is characterized by multiple topics for each paper and hence WARS of all categories are mined.

Shorter WARs (i.e., WARs with few authors and a topic) within all categories are more numerous than longer ones, because they are most likely to satisfy the support threshold. However, as discussed in Section 4.1, long WARs provide interesting information about large research groups, which cannot be easily inferred from the Author-Topic Model (Rosen-Zvi et al, 2012).

4.2.3 Impact of the minimum weighted support threshold

Figures 10 and 11 show the cumulative distribution of the number of A-T WARs (chosen as representative) mined from the *Disorder* and *ATM* datasets, respectively, by varying the value of the weighted support threshold. The plots were generated by counting the number of A-T WARs for each distinct value of wsup while keeping the value of wconf fixed to its standard value (50%).

As expected, the number of mined WARs decreases while considering higher wsup values.

4.2.4 Impact of the minimum weighted confidence threshold

We analyzed also the effect of the weighted confidence threshold on the cardinality of the mined WARs. Figures 12 and 13 show the cumulative distribution of the number of A-T WARs (chosen as representative) mined from the *Disorder* and *ATM* datasets, respectively, by varying the value of the weighted confidence threshold. The plots were generated by counting the number of A-T

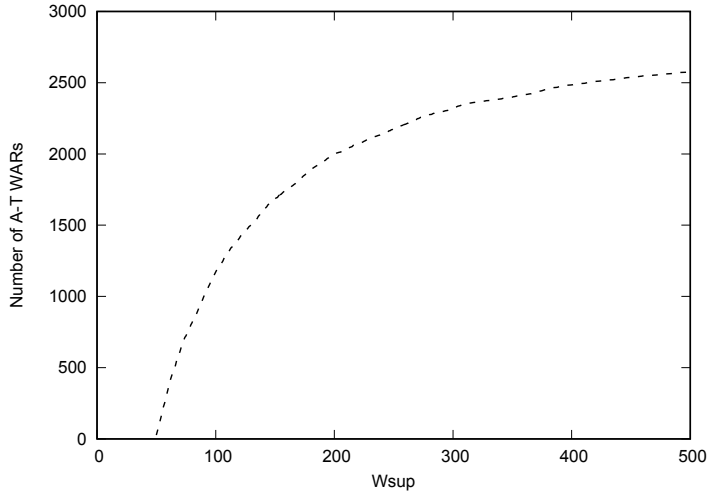

Fig. 10 *Disorder dataset*: Cumulative A-T WAR distribution w.r.t. wsup. wconf=50%.

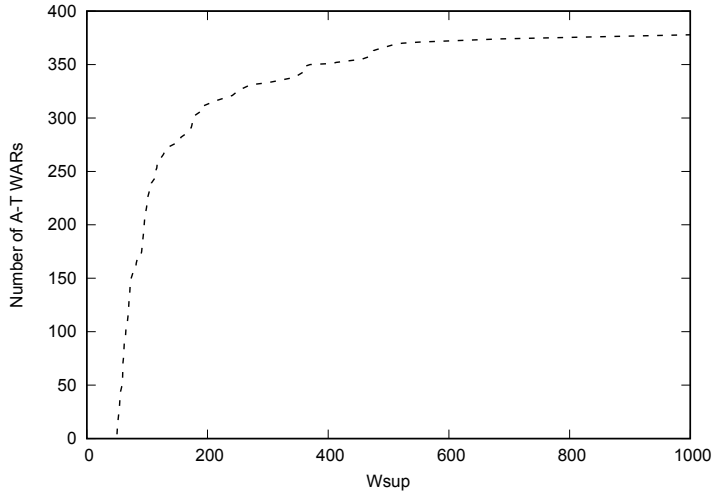

Fig. 11 *ATM dataset*: Cumulative A-T WAR distribution w.r.t. wsup. wconf=50%.

WARs for each distinct value of wconf while keeping the value of wsup fixed to its standard value (50).

The results show that the confidence threshold is not very selective, because most of the mined WARs have fairly high confidence (less than 20% of the WARs have wconf below 80%). The reason is that most groups of researchers have produced highly influential works on a single topic. Thus, the confidence of the corresponding rule is very high. Conversely, the confidence of A-T WARs decreases in case a group has produced scientific works on many different topics. Notice that since publications are weighted by the correspond-

ing number of citations, the collaborations that did not produce any influential works are automatically penalized.

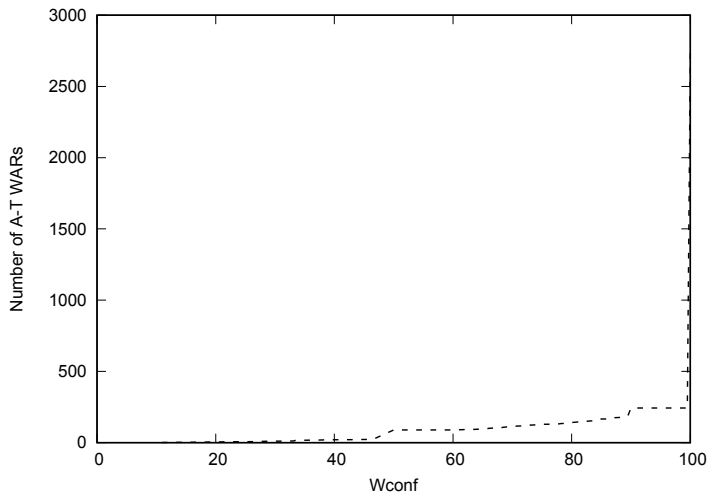

Fig. 12 *Disorder dataset*: Cumulative A-T WAR distribution w.r.t. wconf. wsup=50.

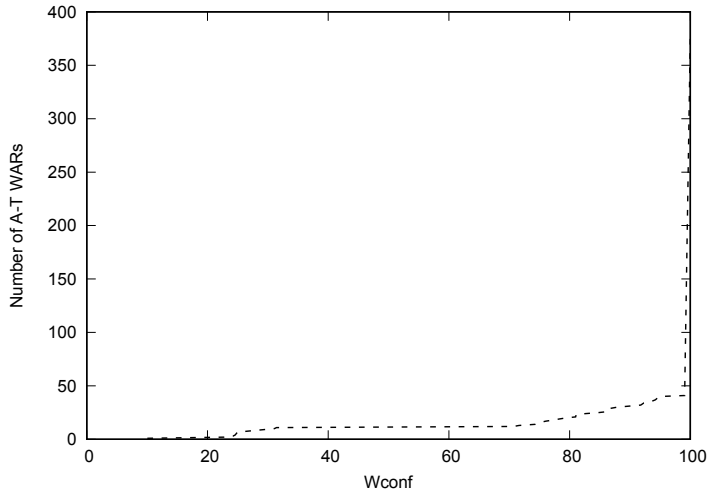

Fig. 13 *ATM dataset*: Cumulative A-T WAR distribution w.r.t. wconf. wsup=50.

4.2.5 Complexity and execution time

We experimentally analyzed the execution time spent by our approach on *ATM* and *Disorder* datasets.

The most computationally intensive tasks are (i) ATM topic detection and (ii) WAR mining. Data preparation and WAR ranking have negligible impact of the execution time. The time complexity of ATM topic detection is of order of the total number of word tokens in the analyzed dataset multiplied by the number of topics. On the *ATM* dataset each run of the generative process takes approximately 20s. This step is not needed on *Disorder* as topics were directly extracted from publication metadata.

The WAR mining process has linear complexity with respect to the number of mined (closed) itemsets, which, in turn, is combinatorial with the number of items ($2^{\#items}$ in the worse case) (Han et al, 2000). Therefore, the time complexity is super-linear with the number of word tokens in the publication documents. For example, on the *Disorder* dataset the WAR mining process took approximately 35s with wsup=100 (approximately 3700 mined WARs), 238s with wsup=50 (21000 mined WARs), and 998s with wsup=25 (23200 mined WARs).

We compared also the performance of the WAR mining process based on closed itemsets with that of a variant of the original process based on all the frequent itemesets (including non-closed itemsets). By relaxing the constraint on closed itemset mining, more than 100 millions of frequent itemsets were generated from both the Disorder and ATM datasets by enforcing a relatively high wsup value (100). The number of the mined frequent itemsets is at least three orders of magnitude larger than those of closed itemsets. The rule generation process on top of frequent itemsets did not terminate due to the huge number of candidate rule combinations (7GB of itemsets for the Disorder dataset, more than 15 GB for the ATM dataset). Therefore, the WAR mining and exploration process becomes practically unfeasible. The reason is that since many articles have a large number of authors, extracting all the frequent itemsets would generate a huge number of redundant patterns. Conversely, closed itemsets represent a more compact representation of the data recurrences.

5 Conclusions and future work

This paper addresses the problem of discovering and ranking fruitful cross-topic collaborations among researchers. The aim is to characterize each research collaboration by discovering the main topics covered and their relative importance in terms of attention given by the research community. To address this issue, it proposes a data mining-oriented methodology, which relies on weighted association rule-based techniques. Weighted Association Rules are interpretable patterns which provide valuable insights into publication data.

The experiments, which were conducted on PubMed (NCBI, 2017) and OMIM (Hamosh et al, 2000) databases, highlight cross-topic collaborations among multiple authors which cannot be easily inferred using traditional models (e.g., the ATM by (Rosen-Zvi et al, 2012)).

As future work, we plan to investigate two complementary research directions, which can be summarized by the following open research issues:
(i) *What are the most appropriate objective and subjective measures to evaluate the interestingness of a rule? How can we effectively drive the user exploration of the mined rules?*
(ii) *Is the proposed methodology applicable to enhance the quality of the peer reviewing process of academic paper?*

To answer to the research question (i), we envision the integration in the proposed methodology of more advanced rule quality indices, both objective (e.g., growth rate (Dong and Li, 1999), chi square (Silverstein et al, 1998), confidence constraint (Baralis et al, 2012)) and subjective (e.g., (Liu et al, 2000)). To this purpose, we plan to carry on a thorough experimental analysis of the impact of different rule quality metrics (Tan et al, 2002) on the quality of the exploratory rule-based models. Furthermore, we aim at collecting the user relevance feedbacks on the mined rules by enriching the Web-based interface available at http://dbdmg.polito.it/CSCA. These feedback scores can be exploited to enhance the quality of the generated model or to refine the process of rule generation based on users' preferences.

To address the research issue (ii), we aim at exploiting the mined WARs to solve the Reviewer Assignment Problem (Kou et al, 2015a,b). Specifically, in the peer reviewing process academic papers are assigned to anonymous reviewers with complementary expertise to assess the innovative contribution of their submitted work. To support journal editors in reviewer assignment, we plan to first extract the ATM topics from the publication records of each candidate reviewer and then discover WARs representing reliable associations between reviewers and topics. The mined WARs can be exploited either to drive the assignment of new reviewers to paper under review or to identify potential conflicts in past reviewer assignments (i.e., reviewers who frequently wrote together assigned to the same paper).

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th VLDB conference, pp 487–499

Agrawal R, Imielinski T, Swami (1993) Mining association rules between sets of items in large databases. In: ACM SIGMOD 1993, pp 207–216

Baralis E, Cagliero L, Cerquitelli T, Garza P (2012) Generalized association rule mining with constraints. Inf Sci 194:68–84, DOI 10.1016/j.ins.2011.05.016

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022, URL http://dl.acm.org/citation.cfm?id=944919.944937

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Seventh International World-Wide Web Conference (WWW 1998), URL http://ilpubs.stanford.edu:8090/361/

Cagliero L, Garza P (2014) Infrequent weighted itemset mining using frequent pattern growth. IEEE Trans Knowl Data Eng 26(4):903–915, DOI 10.1109/TKDE.2013.69, URL http://dx.doi.org/10.1109/TKDE.2013.69

Cagliero L, Garza P, Kavoosifar MR, Baralis E (2017) Identifying collaborations among researchers: a pattern-based approach. In: Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017) co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), Tokyo, Japan, August 11, 2017., pp 56–68, URL http://ceur-ws.org/Vol-1888/paper5.pdf

Ding Y, Zhang G, Chambers T, Song M, Wang X, Zhai C (2014) Content-based citation analysis: The next generation of citation analysis. JASIST 65:1820–1833

Dong G, Li J (1999) Efficient mining of emerging patterns: Discovering trends and differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '99, pp 43–52, DOI 10.1145/312129.312191, URL http://doi.acm.org/10.1145/312129.312191

Hamosh A, Scott A, Amberger J, Valle D, McKusick V (2000) Online mendelian inheritance in man (omim). Human Mutation 15(1):57–61, DOI 10.1002/(SICI)1098-1004(200001)15:1¡57::AID-HUMU12¿3.0.CO;2-G

Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. In SIGMOD'00, Dallas, TX

Hirsch JE (2010) An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. Scientometrics 85(3):741–754, DOI 10.1007/s11192-010-0193-9, URL http://dx.doi.org/10.1007/s11192-010-0193-9

Kim HJ, An J, Jeong YK, Song M (2016) Exploring the leading authors and journals in major topics by citation sentences and topic modeling. In: BIRNDL@JCDL

Kou NM, U LH, Mamoulis N, Gong Z (2015a) Weighted coverage based reviewer assignment. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, SIGMOD '15, pp 2031–2046, DOI 10.1145/2723372.2723727, URL http://doi.acm.org/10.1145/2723372.2723727

Kou NM, U LH, Mamoulis N, Li Y, Li Y, Gong Z (2015b) A topic-based reviewer assignment system. Proc VLDB Endow 8(12):1852–1855, DOI 10.14778/2824032.2824084, URL http://dx.doi.org/10.14778/2824032.2824084

Li B, Hou YT (2016) The new automated ieee infocom review assignment system. IEEE Network 30(5):18–24, DOI 10.1109/MNET.2016.7579022

Liu B, Hsu W, Chen S, Ma Y (2000) Analyzing the subjective interestingness of association rules. IEEE Intelligent Systems and their Applications 15(5):47–

55, DOI 10.1109/5254.889106

Loper E, Bird S (2002) NLTK: the Natural Language Toolkit. In: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, ETMTNLP '02, pp 63–70, DOI 10.3115/1118108.1118117, URL http://dx.doi.org/10.3115/1118108.1118117

Lu C, Zhang C, Ma S (2015) How does citing behavior for a scientific article change over time?: A preliminary study. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, American Society for Information Science, Silver Springs, MD, USA, ASIST '15, pp 97:1–97:4, URL http://dl.acm.org/citation.cfm?id=2857070.2857167

Mutschke P (2003) Mining Networks and Central Entities in Digital Libraries. A Graph Theoretic Approach Applied to Co-author Networks, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 155–166. DOI 10.1007/978-3-540-45231-7_15, URL https://doi.org/10.1007/978-3-540-45231-7_15

NCBI (2017) National Center for Biotechnology Information Website. Available at http://www.ncbi.nlm.nih.gov/ Last access: May 2017. URL http://www.ncbi.nlm.nih.gov/

Newman MEJ (2001) Scientific collaboration networks. i. network construction and fundamental results. Rev E 64:2001

Rosen-Zvi M, Griffiths TL, Steyvers M, Smyth P (2012) The author-topic model for authors and documents. CoRR abs/1207.4169, URL http://arxiv.org/abs/1207.4169

Silverstein C, Brin S, Motwani R (1998) Beyond market baskets: Generalizing association rules to dependence rules. Data Mining and Knowledge Discovery 2(1):39–68, DOI 10.1023/A:1009713703947, URL https://doi.org/10.1023/A:1009713703947

Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T (2004) Probabilistic author-topic models for information discovery. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '04, pp 306–315, DOI 10.1145/1014052.1014087, URL http://doi.acm.org/10.1145/1014052.1014087

Sun K, Bai F (2008) Mining weighted association rules without preassigned weights. IEEE Transactions on Knowledge and Data Engineering 20(4):489–495

Tan PN, Kumar V, Srivastava J (2002) Selecting the right interestingness measure for association patterns. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '02, pp 32–41, DOI 10.1145/775047.775053, URL http://doi.acm.org/10.1145/775047.775053

Tan PN, Steinbach M, Kumar V (2005) Introduction to Data Mining. Addison-Wesley

Tang J, Zhang J, Yao L, Li JZ, Zhang L, Su Z (2008) Arnetminer: extraction and mining of academic social networks. In: KDD

Tao F, Murtagh F, Farid M (2003) Weighted association rule mining using weighted support and significance framework. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'03, pp 661–666

Waltman L, van Eck NJ (2015) Field-normalized citation impact indicators and the choice of an appropriate counting method. Journal of Informetrics 9(4):872 – 894, DOI https://doi.org/10.1016/j.joi.2015.08.001, URL http://www.sciencedirect.com/science/article/pii/S1751157715300456

Wang J, Han J, Pei J (2003) Closet+: searching for the best strategies for mining frequent closed itemsets. In: Getoor L, Senator TE, Domingos P, Faloutsos C (eds) Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 236–245

Wang W, Yang J, Yu PS (2000) Efficient mining of weighted association rules (WAR). In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'00, pp 270–274

White S, Smyth P (2003) Algorithms for estimating relative importance in networks. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '03, pp 266–275, DOI 10.1145/956750.956782, URL http://doi.acm.org/10.1145/956750.956782

Zhang G, Ding Y, Milojevic S (2013) Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. JASIST 64:1490–1503