Identifying collaborations among researchers: a pattern-based approach

Identifying collaborations among researchers: a pattern-based approach

Luca Cagliero, Paolo Garza, Mohammad Reza Kavoosifar, and Elena Baralis

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

Abstract. In recent years a huge amount of publications and scientific reports has become available through digital libraries and online databases. Digital libraries commonly provide advanced search interfaces, through which researchers can find and explore the most related scientific studies. Even though the publications of a single author can be easily retrieved and explored, understanding how authors have collaborated with each other on specific research topics and to what extent their collaboration have been fruitful is, in general, a challenging task.
This paper proposes a new pattern-based approach to analyzing the correlations among the authors of most influential research studies. To this purpose, it analyzes publication data retrieved from digital libraries and online databases by means of an itemset-based data mining algorithm. It automatically extracts patterns representing the most relevant collaborations among authors on specific research topics. Patterns are evaluated and ranked according to the number of citations received by the corresponding publications.
The proposed approach was validated in a real case study, i.e., the analysis of scientific literature on genomics. Specifically, we first analyzed scientific studies on genomics acquired from the OMIM database to discover correlations between authors and genes or genetic disorders. Then, the reliability of the discovered patterns was assessed using the PubMed search engine. The results show that, for the majority of the mined patterns, the most influential (top ranked) studies retrieved by performing author-driven PubMed queries range over the same gene/genetic disorder indicated by the top ranked pattern.

Keywords. Weighted itemset mining, data mining and knowledge discovery

1 Introduction

Plenty of scientific studies have been published on scientific journal, books, and conference proceedings. To deepen their knowledge on specific research topics researchers commonly explore influential studies written by the most renowned authors. Digital libraries and online databases (e.g., PubMed [12], OMIM [4]) play a fundamental role in supporting researchers in their studies. For example, topic- and author-driven searches are supported by most of the renowned digital

libraries. Typically, searches are manually performed to retrieve the publications of interest.

In literature many studies have analyzed the correlation between the authors of scientific papers and the topics covered by the scientific literature (e.g., [3, 7, 15, 19]). For example, existing approaches allow us to identify the most relevant publications of an author on a given topic. An established way to measure the relevance of a publication in the research community is to count the number of received citations [11]. Hence, a large body of work addressed the problem of identifying most influential research works covering a given topic by means of citation content analysis [3, 7, 19]. Citations not only indicate the relevance of a publication but can be exploited also to assess the influence/reputation of individual researchers in their community [13].

Since in several domains research works are often fruit of joint efforts of many researchers, a parallel research issue is the study of the effectiveness of the research collaborations among multiple authors. Manually identifying the most fruitful collaborations on a given research topic is, in general, a challenging task, because it requires correlating the contribution of multiple authors on a specific topic by evaluating the significance of their joint research studies with respect to the existing literature. Hence, automated solutions aimed at analyzing publication data and automatically discovering fruitful research collaborations would be desirable.

To address the aforesaid issue, we propose an automated data mining strategy based on the analysis of publication data acquired from digital libraries and online databases. The aim is to identify interesting correlations between the authors of most influential research studies on specific research topics. To analyze publication data we apply a variant of an FP-Growth-like weighted itemset mining algorithm [2]. Weighted itemset mining is an established data mining technique that focuses on discovering recurrent combinations of items, characterized by different importance levels, from transactional data [14, 16, 18]. In our context, items represent authors or research topics. We discover a new type of pattern, namely the *Authors-Topic Pattern* (ATP), which represents combinations of authors and topics that frequently co-occur in the analyzed data (i.e., they are associated with a large number of publications). To consider also the impact of the research collaboration on the research community each publication in the source data is enriched with the current number of received citations. Then, pattern relevance is measured as a weighted frequency of occurrence in the analyzed data (hereafter denoted as *influence*). To pinpoint for each collaboration among multiple authors the research topics that have produced most authoritative studies, the corresponding patterns are ranked by decreasing influence. Since the extracted patterns are easily interpretable, users may easily explore the top ranked patterns generated by the automatic data mining process.

We experimentally evaluated the effectiveness of the proposed methodology in a real case study, i.e., the analysis of scientific studies on genomics and genetics acquired from two independent libraries (i.e., OMIM [4] and PubMed [12]). In the context under analysis, the ATPs mined from OMIM data represent com-

binations of researchers working together on a specific genetic disorder or gene. We assessed the reliability of the discovered patterns by exploiting the search engine of the PubMed library. Specifically, for each ATP mined from OMIM data we performed an author-driven query on PubMed to find the most related publications co-authored by the same researchers indicated in the pattern. The query returned a ranked list of related publications. The results show that, for the majority of the top ranked (automatically generated) ATPs, the most influential (top ranked) studies returned by author-driven PubMed queries range over the same gene/genetic disorder indicated by the pattern. Hence, ATPs compactly represent salient information about research collaborations on genomics. The manual retrieval of the same information would entail performing many PubMed search queries and then combining the results, which can be a non-trivial and potentially time-consuming task.

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

Studying the impact of scientists' research based on the citations received by their academic publications is a known research problem (e.g., [3, 7, 19]). Citation content analysis is a common way to tackle this problem. 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, in [7] the authors analyzed the sentences including citation expressions to identify interesting characteristics of scholarly communication. The works presented in [3, 19] classified citations based on their semantics to gain insights into the relationships between authors and topics. A social network of academic researchers has been proposed in [15]. The authors automatically extracted researcher profiles from the Web and integrated publication data into the network from existing digital libraries. The proposed academic search system, called ArnetMiner, adopted a predefined Author-Topic model to relate major research topics to most influential authors. However, to the best of our knowledge, it can not be trivially adopted to automatically identify fruitful collaborations among multiple authors on a specific topic. Therefore, the goal of this work is complementary to the above-mentioned approaches.

To assist editors in the peer review of scientific papers a significant research effort has been devoted to proposing new methodologies for matching paper topics with researchers' expertise. For example, in [8–10] the authors addressed the problem of choosing a pool of reviewers for a given article based on the expertise of a potentially large set of candidate reviewers and on the main topics covered by the paper under review. They 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 major topics of the paper and each candidate reviewer has a reasonable number of reviews to do. Conversely, the problem addressed in this paper is not an optimization issue. Even though discovering fruitful collaborations among authors of scientific studies may help conference chairs and journal editors to effectively plan reviewer assignments, the patterns extracted by our methodology are general and they have not been specifically designed to address the reviewer assignment problem.

A parallel research effort has been devoted to efficiently extracting itemsets and association rules from weighted data [2, 14, 18, 16]. This problem extends the traditional association rule mining task, which was first introduced in [1] 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. In [18] the authors proposed to extract weighted association rules, i.e., rules including weights denoting item significance are extracted. In [16] and [2] weights are used to drive the frequent and infrequent itemset mining processes, respectively, while in [14] weights are automatically generated by means of graph indexing techniques. This paper applies a variant of a weighted itemset mining algorithm [2] to discover a new type of patterns, which represents significant correlations between authors and research topics.

3 Scientific Collaboration Analyzer

Scientific Collaboration Analyzer (SCA) is a new data mining-oriented methodology to analyze the scientific literature accessible through digital libraries and online database. The goal is to identify sets of authors (of arbitrary size) whose joint research on specific topics has produced publications with significantly high impact. The methodology consists of two main steps: (i) *Data collection and preparation.* Publication data and citations are acquired from online sources, collected into a unique repository and tailored to the next mining process (see Section 3.1). (ii) *Pattern discovery, evaluation, and ranking.* Patterns that represent combinations of authors and topics are extracted from the prepared data and ranked according to ad hoc evaluation metrics (see Section 3.2). A more thorough description of each step follows.

3.1 Data preparation

Publication data are acquired from digital libraries and online databases and stored in a unique repository. For our purposes, for each publication we acquire the following information: (i) the Digital Object Identifier (DOI) of the publication, (ii) the list of authors, (iii) the list of the research topics that are mainly discussed in the publication, and (iv) the number of citations received.

Table 1. Example of weighted transactional dataset

Pub. id	#cit.	Authors	Topics
1	10	$(Author : Brown,\ J.)$, $(Author : Smith,\ L.)$	$(Topic : A)$, $(Topic : X)$
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	$(Author : Smith,\ L.)$	$(Topic : Z)$

Table 2. Patterns ranked by decreasing influence. Minimum influence threshold $mininf$ =20

Authors-Topic Pattern	Influence
$\{(Author : Brown,\ J.)$,$(Author : Smith,\ L.)$, $(Topic : X)\}$	25
$\{(Author : Smith,\ L.)$,$(Topic : Z)\}$	23

The current number of citations is considered because it has largely been used to assess the influence/popularity of a publication in the research community [11]. However, since the proposed methodology is general, different measures can be easily integrated as well (e.g., the Hirsch index [6]).

To allow pattern mining, publication data are collected 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 number of citations received by the corresponding publication. We consider as transaction weight, denoting the relevance of each publication, the number of citations.

Transactions consist of sets of items, where items are publication authors (e.g., *Smith, L.*), or research topics (e.g., *topic X*). 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 the number of citations received by 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 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 *# cit.*). 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.

3.2 Pattern discovery, evaluation, and ranking

This step entails discovering a new type of pattern from the prepared weighted transactional dataset, namely the Authors-Topic Pattern (ADP). It represents a potentially interesting correlation between a set of authors and a research topic.

Preliminaries. Pattern extraction relies on itemset mining techniques. Frequent itemset mining [1] is an established data mining technique to discover recurrent correlations among data items hidden in large datasets. 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 [1]. 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 [17], can be extracted. An itemset is *closed* if there exists no superset that has the same support as this original itemset.

Pattern definition. For our purposes, we are interesting in mining a particular type of closed itemsets: the combinations of authors with a specific research topic. Hereafter, we will denote it as Authors-Topic Patterns (ATP).

Definition 3. Authors-Topic Patterns (ATP). *Let I be a closed k-itemset. I is an ATP if (i) it contains one or more items feature:v_q such that feature=Author, and (ii) it contains exactly one item such that feature=Topic.*

Recalling the running example, Table 2 reports some examples of ATPs mined from the dataset in Table 1. For example, $\{(Author : Brown,\ J.),(Author : Smith,\ L.), (Topic : X)\}$ indicates that researchers Brown J. and Smith L. have co-authored many research publications on topic X.

Pattern evaluation. The support quality index of an itemset [1] does not consider the relative importance of each transaction in the source dataset. 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 topic that have achieved a high impact, we extended the standard itemset mining problem by integrating item weights [18]. Specifically, item occurrences within each transaction (publication) are weighted by the corresponding number of citations. 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 *influence* of an ATP as a weighted frequency of occurrence of the itemset in the weighted transactional dataset.

Definition 4. ATP influence. *Let $\mathcal{D}$ be a weighted transactional dataset and I be an ATP. Let tw_j: $\langle t_j,\ C(p_j)\rangle$ be an arbitrary weighted transaction in $\mathcal{D}$. The influence of ATP I in $\mathcal{D}$, hereafter denoted by inf(I), is defined as follows:*

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

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

Pattern filtering and ranking. To filter out less interesting patterns a minimum influence threshold is enforced. Specifically, given a weighted transactional dataset and a user-specified minimum influence threshold $mininf$, we extract all the ATPs whose influence value is above or equal to the threshold. Patterns can be sorted by decreasing influence to quickly retrieve the most relevant ones.

3.3 The extraction algorithm

Many frequent weighted itemset mining algorithms have already been proposed in literature (e.g., [2, 14, 16, 18]). To accomplish the ATP mining task from

weighted transactional data, we adopt a variant of the algorithm first proposed in [2], which performs FP-Growth-like closed itemset mining. FP-Growth [5] relies on an FP-tree data model, i.e., a compact, tree-based representation of the original dataset residing in main memory. To efficiently generate ATPs on top of closed itemsets, we separately extracted closed itemsets for each topic by recursively visiting the FP-tree structure in a depth-first manner.

4 Case study

We investigated the applicability of the proposed methodology in a real case study, i.e., the analysis of the research collaborations on genomics and genetics. To perform our experiments we analyzed publication data that were acquired from the Online Mendelian Inheritance in Man (OMIM) catalog of genetic disorders [4].

The Online Mendelian Inheritance in Man (OMIM) database [4] is one of the most comprehensive and authoritative compendia of human genes and genetic phenotypes. OMIM is part of the National Center for Biotechnology Information (NCBI) system of databases [12] 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) the set of genes correlated with the disorder, (ii) a list of scientific publications ranging over the disorder (for each publication the complete bibliographic information is known), (iii) a textual description of the disorder including references, and (iv) links to other genetics resources.

Our study is focused on discovering from OMIM data sets of researchers that have conducted influential studies on genomics or genetics. To tailor OMIM data to our context of analysis, we considered as topic categories the genetic disorders and the genes discussed in each publication. Specifically, the weighted transactional datasets contains items belonging to three different features: *Author*, *Gene*, and *Genetic Disorder*. To crawl data from the online OMIM database, we exploited the exposed APIs [4]. Instead, to retrieve the number of citations received by each publication we exploited the APIs of the PubMed digital library [12]. The integrated dataset, which were obtained by integrating publication data from OMIM and citation data from PubMed, contains 8825 articles, 34555 authors, 302 disorders, and 1076 genes. The experiments were performed on a 2.67 GHz Intel Xeon workstation with 32 GB of RAM, running Ubuntu Linux 12.04 LTS. The data crawler and the data preparation steps are written in Java, while the pattern mining algorithm is written in C.

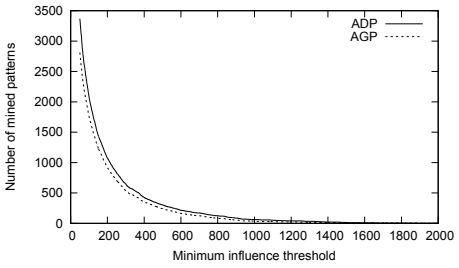

Fig. 1. Number of mined patterns

Table 3. Number of mined patterns vs number of authors in the pattern (minimum influence = 50)

num. of co-authors	num. of ADPs	num. of AGPs
2	760	164
3	447	375
4	286	496
5	230	411
6	151	319
7	117	340
8	101	254
9	82	194
10	69	160

4.1 Pattern characteristics

We analyzed the characteristics of the patterns mined by setting different values of minimum influence threshold (i.e., the minimum number of received citations). Figure 1 plots the number of mined ATPs related to genes, hereafter denotes as Author-Gene Patterns (AGPs) for the sake of brevity, and the number of mined ATPs related to disorders, denoted as Author-Disorder Patterns (ADPs), by varying the $mininf$ threshold value.

By setting the influence threshold we may discover research collaborations whose production has had a rather different impact in their community. For example, by setting $mininf$ to 400 ATPs represent research teams whose scientific studies on a specific gene/disorder have produced at least 400 citations. By increasing $mininf$ the constraint on the minimum number of citations becomes more selective. Hence, as expected, the number of mined patterns decreases more than linearly while increasing the $mininf$ value. The two curves (those related to AGPs and to ADPs) show similar trends.

In Table 3 we categorized the extracted ADPs and AGPs according to the number of authors appearing in each pattern. The reported categorization approximately indicates the average impact of the research groups with a given size. Notably, the influence is not proportional to the number of authors. Small- and medium-size groups (e.g., from 3 to 5 persons) with high research influence are quite frequent. However, a significant number of larger groups (7-8 persons)

have produced influential studies as well[1]. Therefore, it is worth analyzing the correlations between many authors in the considered case study.

4.2 Pattern analysis

We empirically analyzed the strength of the correlations between the research teams and the topic identified by the pattern. For each topic we considered the top-5 ranked patterns, mined from OMIM data, in order of decreasing influence. Each of the selected patterns indicates the most important topic addressed by the research team. The research questions we would like to address in this section are the following:
(A) Are the research team and the topic really correlated with each other?
(B) Among the topics addressed by the team, is the topic indicated in the pattern the most influential one?

To address the above research questions, we assessed the quality of the mined patterns by performing author-driven queries on the PubMed digital library. Specifically, for each pattern we picked the research team (i.e., the author names occurring in the pattern) and we performed author-driven query on PubMed. The PubMed query returns a list of publications ranked by decreasing relevance. If a publication in the top-3 PubMed ranking covers the topic we can conclude that research team and topic are, to some extent, correlated with each other (question (A)). If the publication covering the topic is at the top of the PubMed ranking, we can conclude that the addressed topic is the most influential among those addressed by the research team (question (B)).

We performed the above-mentioned comparison separately for genes and genetic disorders. Specifically, Table 4 summarizes the results of the manual validation performed on the top-5 Authors-Disorder Patterns (i.e., the ATPs related to genetic disorders), while Table 5 reports similar results for the top-5 Authors-Gene Patterns (i.e., the ATPs related to gene). For each pattern we reported the title and the Digital Object Identifier of the top ranked publication returned by PubMed. In most cases, the selected publication matches the topic indicated by the pattern. To show the correlation between genes/disorders and publication content, in Column *PubMed publication title* we highlighted the title keywords that recall, to some extent, the gene name or the genetic disorder indicated in the pattern. For example, the top ranked Author-Gene Pattern in Table 5 concerns gene AUTS1, whose mutations are strongly correlated with the autism disorder. In the top ranked publication returned by PubMed the title made explicit reference to the correlated disorder (*Recurrent de novo mutations implicate novel genes underlying simplex autism risk.*).

5 Conclusion and future work

This paper presents an itemset-based approach to analyzing publication data and to discovering fruitful collaboration among researchers. The proposed method-

[1] Note that genomic and genetic studies are likely to be co-authored by many researchers.

Table 4. Example of Authors-Disorder Patterns

Authors-Disorder Pattern (influence)	**PubMed publication title**	**PubMed publication DOI**
{Author:Siddique T., Author:Deng H. X., Disease:AMYOTROPHIC LATERAL SCLEROSIS} (*inf*=1828)	Inclusions in frontotemporal lobar degeneration with TDP-43 proteinopathy (FTLD-TDP) and **amyotrophic lateral sclerosis (ALS)**, but not FTLD with FUS proteinopathy (FTLD-FUS), have properties of amyloid.	10.1007/s00401-013-1089-6
{Author:Rioux J. D., Author:Silverberg M. S., Disease:INFLAMMATORY BOWEL DISEASE (CROHN DISEASE)} (*inf*=1470)	Host-microbe interactions have shaped the genetic architecture of **inflammatory bowel disease**.	10.1038/nature11582
{Author:Flaherty K. T., Author:Ribas A., Author:Chapman P. B., Disease=MELANOMA CUTANEOUS MALIGNANT SUSCEPTIBILITY TO} (*inf*=1253)	Improved survival with vemurafenib in **melanoma** with BRAF V600E mutation.	10.1056/NEJMoa1103782
{Author:Hamilton S. R., Author:de la Chapelle A., Disease=LYNCH SYNDROME I} (*inf*=1116)	Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (**Lynch syndrome**) and microsatellite instability.	Not available
{Author:Thornton C. A., Author:Swanson M. S., Disease=MYOTONIC DYSTROPHY} (*inf*=889)	Splicing biomarkers of disease severity in **myotonic dystrophy**.	10.1002/ana.23992

Table 5. Example of Authors-Gene Patterns

Authors-Gene Pattern (influence)	**PubMed publication title**	**PubMed publication DOI**
{Author:O'Roak B. J., Author:Vives L., GeneSymbols:AUTS1} (*inf*=1369)	Recurrent de novo mutations implicate novel genes underlying simplex **autism** risk.	10.1038/ncomms6595
{Author:Spiegel A. M., Author:Marx S. J., GP-GeneSymbols:MEN1} (*inf*=620)	Recapitulation of pancreatic neuroendocrine tumors in human multiple endocrine **neoplasia** type I syndrome via Pdx1-directed inactivation of Men1.	10.1158/0008-5472.CAN-08-3662
{Author:Allen R. P., Author:Earley C. J., GP-GeneSymbols:RLS1} (*inf*=219)	Intervening **Leg Movements Disrupt** PLMS Sequences.	10.1093/sleep/zsw023
{Author:Berson E. L., Author:Dryja T. P., GP-GeneSymbols:RP} (*inf*=553)	Novel mutations in the long isoform of the USH2A gene in patients with Usher syndrome type II or non-syndromic **retinitis pigmentosa**.	10.1136/jmg.2009.075143
{Author:Julian B. A., Author:Wyatt R. J., GP-GeneSymbols:IGAN1} (*inf*=311)	GWAS for serum galactose-deficient **IgA1** implicates critical genes of the O-glycosylation pathway.	10.1371/journal.pgen.1006609

ology generates interpretable patterns that compactly represent collaborations among researchers that have produced the most influential studies. The applicability and effectiveness of the proposed methodology has been experimentally evaluated in real case study, i.e., the analysis of the publications related to genomic and genetics studies. We plan to apply our methodology to support reviewer assignment in the process of paper peer reviewer. We aim at integrating Authors-Topic associations into existing optimization-based strategies (e.g., [8, 9]). Considering correlations between multiple authors and a topic would help editors to diversify assignments across researchers with complementary expertise.

References

1. R. Agrawal, T. Imielinski, and Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD 1993*, pages 207–216, 1993.
2. L. Cagliero and P. Garza. Infrequent weighted itemset mining using frequent pattern growth. *IEEE Trans. Knowl. Data Eng.*, 26(4):903–915, 2014.
3. Y. Ding, G. Zhang, T. Chambers, M. Song, X. Wang, and C. Zhai. Content-based citation analysis: The next generation of citation analysis. *JASIST*, 65:1820–1833, 2014.
4. A. Hamosh, A. Scott, J. Amberger, D. Valle, and V. McKusick. Online mendelian inheritance in man (omim). *Human Mutation*, 15(1):57–61, 2000.
5. J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *In SIGMOD'00, Dallas, TX*, May 2000.
6. J. E. Hirsch. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3):741–754, Dec. 2010.
7. H. J. Kim, J. An, Y. K. Jeong, and M. Song. Exploring the leading authors and journals in major topics by citation sentences and topic modeling. In *BIRNDL@JCDL*, 2016.
8. N. M. Kou, L. H. U., N. Mamoulis, and Z. Gong. Weighted coverage based reviewer assignment. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 2031–2046, New York, NY, USA, 2015. ACM.
9. N. M. Kou, L. H. U, N. Mamoulis, Y. Li, Y. Li, and Z. Gong. A topic-based reviewer assignment system. *Proc. VLDB Endow.*, 8(12):1852–1855, Aug. 2015.
10. B. Li and Y. T. Hou. The new automated ieee infocom review assignment system. *IEEE Network*, 30(5):18–24, September 2016.
11. C. Lu, C. Zhang, and S. Ma. 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*, ASIST '15, pages 97:1–97:4, Silver Springs, MD, USA, 2015. American Society for Information Science.
12. NCBI. National Center for Biotechnology Information Website. Available at http://www.ncbi.nlm.nih.gov/ Last access: May 2017, 2017.
13. A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280, 2007.
14. K. Sun and F. Bai. Mining weighted association rules without preassigned weights. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):489 –495, 2008.

15. J. Tang, J. Zhang, L. Yao, J.-Z. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, 2008.
16. F. Tao, F. Murtagh, and M. Farid. 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*, pages 661–666, 2003.
17. J. Wang, J. Han, and J. Pei. Closet+: searching for the best strategies for mining frequent closed itemsets. In L. Getoor, T. E. Senator, P. Domingos, and C. Faloutsos, editors, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 236–245, 2003.
18. W. Wang, J. Yang, and P. S. Yu. Efficient mining of weighted association rules (WAR). In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'00*, pages 270–274, 2000.
19. G. Zhang, Y. Ding, and S. Milojevic. Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *JASIST*, 64:1490–1503, 2013.