Indirect Association Rules Mining in Clinical Texts

Svetla Boytcheva[0000−0002−5542−9168]

Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Bulgaria
svetla.boytcheva@gmail.com

Abstract. This paper presents a method for structured information extraction from patient status description. The proposed approach is based on indirect association rules mining (IARM) in clinical text. This method is language independent and unsupervised, that makes it suitable for applications in low resource languages. For experiments are used data from Bulgarian Diabetes Register. The Register is automatically generated from pseudonymized reimbursement requests (outpatient records) submitted to the Bulgarian National Health Insurance Fund in 2010–2016 for more than 5 million citizens yearly. Experiments were run on data collections with patient status data only. The great variety of possible values (conditions) makes this task challenging. The classical frequent itemsets mining algorithms identify just few frequent pairs only even for small minimal support. The results of the proposed IARM method show that attribute-value pairs of anatomical organs/systems and their condition can be identified automatically. IARM approach allows extraction of indirect relations between item pairs with support below the minimal support.

Keywords: Data Mining · Text Mining · Health Informatics.

1 Motivation

Structured information extraction of patient status is important for many Healthcare Management tasks like diagnosis, treatment effect assessment, patient profiling, etc. Usually outpatient records (ORs) contain detailed description of patient status as free text only. Raw clinical text usually used telegraphic style of the patient status, rather than full sentences. There are also a lot of typo and abbreviations, that makes difficult usage of standard syntax parsers. Without of resources like lexicons and ontologies with medical terminology the solution of task for complex events and relations extractions is almost impossible. Still there is a lack of resources and natural language processing (NLP) tools for non-English clinical texts processing [21] albeit there is an increasing research efforts in this area. There are still non existing translations of SNOMED[1], Medical Sub-

[1] SNOMED, https://www.snomed.org/

ject Headings (MeSH)[2] and Unified Medical Language System (UMLS)[3] for the majority of languages.

Manual annotation of many clinical text documents is time and effort consuming task. The rich vocabulary of the medical terminology is one of the major obstacles for scalability of methods developed and evaluated on the golden standard basis. Thus for NLP of clinical texts there are needed language independent methods that are unsupervised and do not rely on resources. In previous research we shown that some data mining algorithms for frequent patterns mining and frequent sequence mining can be used efficiently to extract structured information for risk factors and some complex relations between diagnosis [5], [4].

The classical frequent itemsets mining and association rules generation algorithms identify just few frequent pairs only even for small minimal support. The infrequent data contain interesting and important information. Indirect association rules mining (IARM) identify relations between items $\{X, Y\}$ that rarely occur together in the same transaction. In case the presence in transaction of both items X and Y depend on presence of some other set M, then it is said that X and Y are *indirectly associated* via M [24]. The set M is called *mediator*. In text documents processing indirectly associated words corresponds to words, used in different context of other word, or synonyms, or antonyms. We propose a method based on IAR mining of clinical text. The aim of this research is to identify attribute-value pairs of patient status descriptions.

The paper is structured as follows: Section 2 briefly overviews the research in the area; Section 3 describes the data collection of clinical text used in the experiments; Section 4 presents the formal presentation of the problem and describes in details the proposed method for mining indirect association rules in clinical text; Section 5 shows experimental results and discusses the method applications; Section 6 contains the conclusion and sketches some plans for future work.

2 Related Work

Recently many data mining approaches were successfully used in NLP tasks [18]. Manimaran and Velmurugan show how Apriory [2] algorithm can be used for text mining with application in medical informatics.

One of the earliest comprehensive studies of IAR algorithms for mining transaction data were presented by Tan and Vipin [24], [22], [23]. They present different applications of the method for text, retail data, stock market data, Web click-stream data.

Chen et al [6] present an algorithm MG-Growth for temporal indirect association that takes into account the lifespan of items. Ha et al [11] demonstrate how IARM approach can be used to find hidden correlation among multimedia semantic concepts.

[2] Medical Subject Headings — MESH, https://www.nlm.nih.gov/mesh/
[3] UMLS, https://www.nlm.nih.gov/research/umls/

Some attempts to define more efficient measures in IARM task was presented. Abdullah et al [1] introduce a new measure Definite Factor to indicate the degree of certainty of association rules. Hamano and Sato [12] define a measure for mining Negative Association rules effectively without domain knowledge. Wan and An [26] describe HI-mine - an efficient algorithm, based on a novel data structure, called HIstruct. Kazienko [17] proposes IDARM* Algorithm that uses pre-calculated direct rules for complete IARM. He presents application of the proposed algorithm for recommendation system for web pages.

There are many applications of IARM in the domain of medical informatics. Tsuruoka et al [25] present a real-time text mining system FACTA+ for finding and visualizing indirect associations between biomedical concepts from MEDLINE abstracts. Another interesting application of IARM in medical informatics is presented by Kang and Wagacha [16]. They investigate ICD–9[4] disease diagnosis associations in big collection of Electronic Health Records. Wright et al [27] apply IARM for identifying associations between medications, laboratory results and problems.

3 Materials

Bulgarian National Diabetes Register [3] is automatically generated from a data repository of about 262 million pseudonymized outpatient records (ORs) submitted to the Bulgarian National Health Insurance Fund (NHIF) in period 2010–2016 for more than 5 million citizens yearly. The NHIF collects for reimbursement purpose all ORs produced by General Practitioners and the Specialists from Ambulatory Care for every patient clinical visit.

ORs are stored in the repository as semi-structured files with predefined XML-format. Structured information describe the necessary data for health management like visit date and time; pseudonymized personal data and visit-related information, demographic data (age, gender, and demographic region), etc. All diagnoses are presented by ICD–10[5] codes and the name according to the standard nomenclature. The most important information concerning patient status and case history is provided like free text. ORs contain paragraphs of unstructured text provided as separate XML tags: “Anamnesis” (Disease history), “Status”, “Clinical tests”, and “Prescribed treatment”.

For experiments is used a data collections of ORs from Bulgarian National Diabetes Register. For all experiments are used raw ORs, without any preprocessing due to the lack of resources and annotated corpora. The text style for unstructured information is telegraphic. Usually with no punctuation and a lot of noise (some words are concatenated; there are many typos, syntax errors, etc.). The Bulgarian ORs contain medical terminology both in Latin and Bulgarian. Some of the Latin terminology is also used with Cyrillic transcription. Bulgarian language uses inflections. For some Latin medical terms in addition to the original Latin and Greek suffixes are used also prefixes and suffixes specific

[4] http://icd9.chrisendres.com/
[5] http://apps.who.int/classifications/icd10/browse/2016/en#/

for Bulgarian language. This makes the task for natural language processing of the clinical text in Bulgarian quite challenging.

The ORs are written in telegraphic style with phrases rather than full sentences. Usually the ORs list attribute-value (A-V) pairs - anatomical organ/system and its status/condition. Attribute names contain phrases and abbreviations in Cyrillic and Latin. Values can be long descriptions in case of status complications. The order of A-V pairs can vary and parts of the value descriptions can surround the attributes. It is also possible that some attributes share the same value. Sample configurations are shown below.

$$A_1V_1, ..., A_nV_n|V_1A_1, ..., V_nA_n$$
$$V_1...V_kAV_{k+1}...V_n$$
$$A_1, A_2, ..., A_nV|VA_1, A_2, ..., A_n.$$

4 Methods

4.1 Indirect Association Rules Mining

The vocabulary used in all ORs of a data collection S will be called *items* $V = \{v_1, v_2, ..., v_n\}$. For the collection S we extract the set of all different ORs $R = \{r_1, r_2, ..., r_N\}$, where $r_i \subseteq V$. This set corresponds to transactions; the associated unique transaction identifiers (*tids*) will be called ***pids*** (patient identifiers). Each patient interaction with a doctor is viewed as a single OR in R.

Preliminary analysis of N-grams in data collections with AntConc tool[6] show that the majority of the identified N-gram candidates are mainly cliché phrases. There are only seldom examples(about 15%) for true positive N-grams.

Therefore we treat documents as bags of words rather than sequences; they are transformed to itemsets with single word occurrences only.

Given a set of pids S, the support of an itemset I is the number of pids in S that contain I. We denote it as $supp(I)$. We define a threshold called *minsup* (minimum support). Frequent itemset (FI) I is one with at least minimum support count, i.e. $supp(I) \geq minsup$. The task of FPM of S is to find all possible frequent itemsets in S.

The following definition for indirect association rules was proposed by Tan and Vipin [24]:

Definition 1. *(Indirect associated pair) An itempair $\{X; Y\}$ is indirectly associated via a mediator set M if the following conditions hold :*

1. $sup(X; Y) < minsup$ (*Itempair Support Condition*)
2. There exists a non-empty set M such that $\forall M_i \in M$:
a) $sup(X; M_i) \geq ts$; $sup(Y; M_i) \geq ts$ (*Mediator Support Condition*).
b) $d(X; M_i) \geq conf$; $d(Y; M_i) \geq conf$ where $d(p; Q)$ is a measure of the dependence between p and Q (*Dependence Condition*).

[6] Laurence, A. AntConc (Version 3.4. 4w)(Computer software). Tokyo, Japan:Waseda University. http://www.laurenceanthony.net/ (2014)

Condition (1) is needed because an indirect association is significant only if there are seldom occurrences of both items in ORs, i.e. negatively correlated. Condition (2a) is needed to guarantee statistical significance of the mediator set M. Condition (2b) is needed to guarantee that only items highly dependent on both X and Y are used to form the mediator set M. Items in M form close neighborhood.

4.2 Automatic Extraction of Patient Status

The workflow of the proposed method for automatic extraction of patient status from clinical text is shown on Fig. 1. The processes are grouped in three main subtasks:

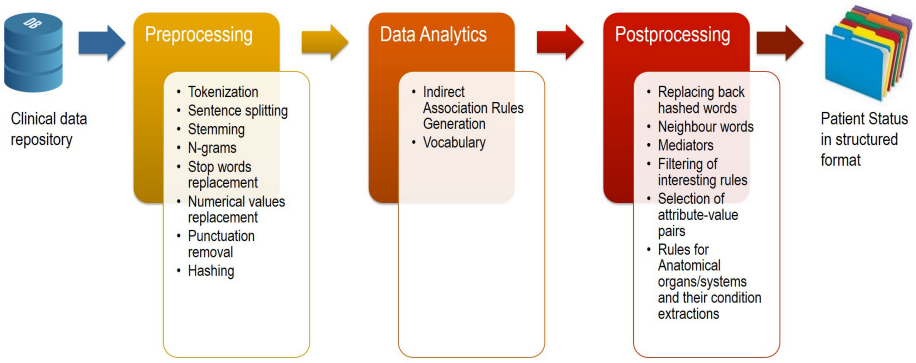

Fig. 1. Workflow

– **Preprocessing** - converts raw ORs data into itemsets. This subtask starts with tokenization of "Status" field only of ORs. The texts in "Status" section is usually short and presented in telegraphic style and without punctuation. Thus the majority of ORs are considered as single sentence only. The next step is stemming that is necessary for reducing inflected words. For this step is used Bulstem [20]. Unfortunately it doesn't works for part of the medical terminology which is in Latin but transliterated in Cyrillic. For next step N-grams identification is used AntConc as we mentioned in the previous subsection. Some of the top ranked N-grams are replaced in the text by single item. After that stop words are identified and removed. ORs contain many numerical values in the "Status" section, like blood pressure, Body mass index, height, weight, etc. All numerical values are replaced by symbol NUM. The punctuation symbols are removed from the text, because they don't matter when ORs are processed as bag of words. Finally itemsets

are generated by applying hashing - replacing each item (word/N-gram) by unique ID and removing duplicates. The itemset is stored in increasing order of the items ID in order to fasten the data mining process.
- **Data Analytics** - applies data mining methods for IARM, FPM and Association Rules (ARs) . For experiments are used Java implementations of the algorithms IndirectRules[24], FPMax [9], and FPGrowthARL [13] from SPMF[7] (Open-Source Data Mining Library) [8].
- **Postprocessing** - packs the result data. The first step is to return back the hashed items in the indirect association rules and frequent itemsets. Presenting results for all indirect pairs and the corresponding mediators. Identification of attributes and their values.

5 Experiments and Results

The associations of the following types are in primary interest of our task for patient status extraction (Fig. 2). In Fig. 2b AX and AY are two attributes and they share mediator set with common values. Such pairs of attributes can be the Bulgarian and Latin terms used for an anatomical organ system.

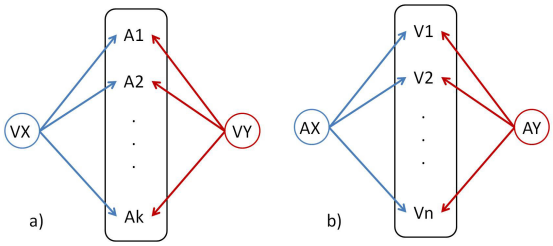

Fig. 2. Attribute-Value indirect association rules

Example 1:
AX=черен дроб и слезка (Bulgarian) (*liver and spleen*)
AY=hepar et lien (Latin) (*liver and spleen*)
M= б.о. (*without complications*), не се палпират увеличени (*do not palpate enlarged*)

Although "liver and spleen" describes a pair of attributes they can not be split further, because there will be violation of the condition (1) of the Definition 1, both indirectly associated items not to occur frequently together in ORs. There is a direct association between them and they can be identified as frequent pair.

Another case for such interesting indirect pair is when an abbreviation and full attribute name are used.

Example 2: For example for Cardiovascular system (CVS) сърдечно-съдова система (CCC):

[7] http://www.philippe-fournier-viger.com/spmf/index.php

AX=CCC (*CVS*)
AY=сърце (*heart*), or
AY=cor (Latin)(*heart*).

In Fig. 2a VX and VY are two values and they share mediator set with common attributes. Usually such values describe general conditions and observations.

Example 3:
VX =добро (*good*)
VY = увредено (*impaired*)
M=общо състояние (*general condition*)

Example 4:
VX =не се палпират увеличени (*do not palpate enlarged*)
VY = увеличени (*enlarged*)
M=лимфни възли (*lymph nodes*) , черен дроб и слезка (*liver and spleen*)

	S00 (General Practitioners)	S05 (Endocrinology)
Patients	10,000	10,000
ORs	123,247	14,753
sentences	791,420	157,448
items (vocabulary)	8,408	2,111
minimal support (minsup)	0.011	0.011
mediator support (ts)	0.7	0.7
minimal confidence (conf)	0.7	0.7
Maximal frequent patterns	81	43
Indirect pairs	195	2,670
Association rules	1,236	7,121

Table 1. Summary of data parameters and results

The experiments were run on 2 collections of ORs from clinical visits to different specialists in Endocrinology (S05), and General Practitioners (S00). Dataset S00 contain 791,420 sentences (transaction), and S05 - 157,448 sentences. Both datasets are sparse with huge number of items (vocabulary) (Fig. 1). Although the values of *minsup* are relatively small, there are only few frequent patterns due to the huge variety of values for each attribute. For most of experiments are used comparable values for minimal support (minsup = 0.015), mediator support (ts =0.7) and minimal confidence (conf = 0.7), i.e. mediator dependency. For performance evaluation are used values for *minsup* in the range [0.015, 0.025] (Fig. 3), and for *conf* are used values in the range [0.2, 0.9] (Fig. 4).

Even small increase in the minsup value (Fig. 3) causes significant drop down of the total number of generated indirect association rules (IAR). Similarly the quantity of generated IAR is exponential decay when minimal confidence increases (Fig. 4).

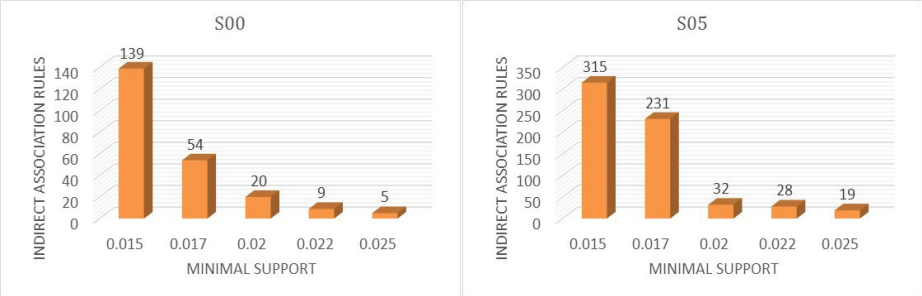

Fig. 3. Minimal support vs Total number of indirect association rules for ts=0.7 and conf=0.7 for datasets S00 and S05

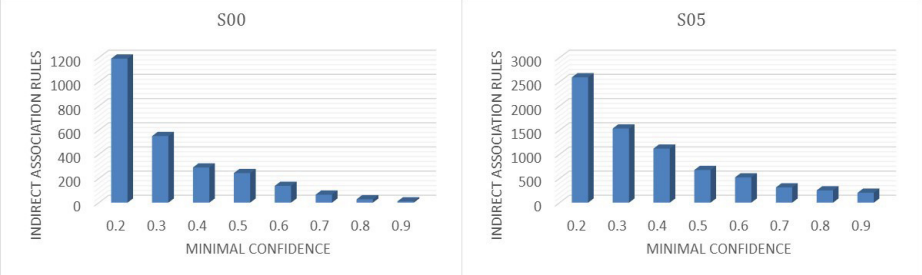

Fig. 4. Minimal confidence vs Total number of indirect association rules for ts=0.7 and minsup=0.015 for datasets S00 and S05

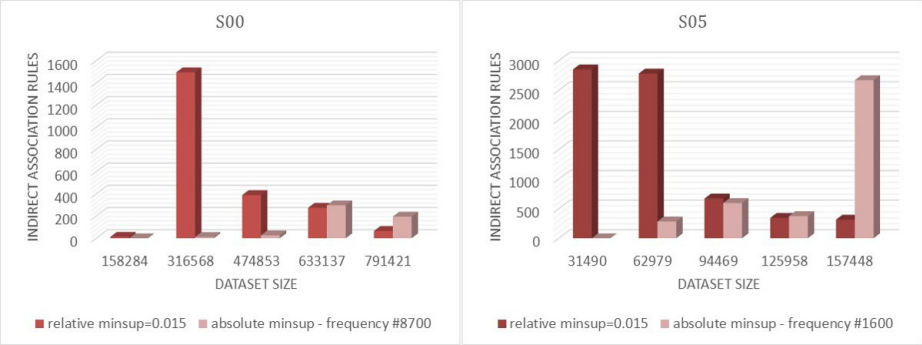

Fig. 5. S00 dataset size vs Total number of indirect association rules for ts=0.7, conf=0.7 for datasets S00 and S05 with different minsup values - absolute and relative

For performance evaluation the datasets S00 and S05 are fractioned on 5 equal subsets (Fig. 5). There are presented results for two experiments - with fixed relative value of minsup, i.e. as percentage of the dataset size. Increasing the dataset size affects the the absolute value of the threshold an it also increases.

Thus the total number of IAR declines. For the second experiment is used same threshold for generating IAR in all datasets. For smaller datasets the IAR have not significant support and the cumulative effect can be obtained for bigger datasets.

Some examples for extracted indirect pairs from S05 are presented below.

```
(X= 1 Y= 7 | mediator= 2 )
sup(X,mediator)= 100  sup(Y,mediator)= 68
conf(X,mediator)= 1.0  conf(Y,mediator)= 0.9685
```

Where the ID 1 = BMI (*Body Mass Index*) and 7 = сч - сърдечена честота (*(heart) frequency*).

The mediator set contains ID 2 = NUM - the symbol by which are replaced all numerical values in preprocessing subtask. In this example X and Y corresponds to different attributes and mediator set contains the type of their value.

The next example presents extracted indirect pair from S00, where X and Y corresponds to different possible values of the attribute presented in the mediator set.

```
(X= 7 Y= 60 | mediator= 9 )
sup(X,mediator)= 101 sup(Y,mediator)= 182
conf(X,mediator)= 0.9439 conf(Y,mediator)= 0.9479
```

The mediator set contains ID 9 = дишане (*breath*) and the ID 7 = отслабено (*weakened*) and 60 = удължено (*prolonged*).

For method accuracy evaluation are used standard metrics Precision Recall and F1-measure, where F1 measure is defined as:

$$F1 = 2 \times \frac{Precission \times Recall}{Precission + Recall}.$$

Experiments are provided by non-exhaustive cross-validation (5 iterations on sets in ratio 4:1 training to test). For S00 are generated 1180 pairs of variable and mediator, and 274 pairs of indirect associated concepts. The size of the test set for S00 is 158,283 sentences. The precision is 0.85, recall is relatively low - 0.33 due to the small number of IAR and huge number of sentences in test set. This dataset is based on ORs written by General practitioners, thus the diversity of attribute-values pairs is higher. The overall evaluation for S00 is F1=0.49. For dataset S05 - 74 pairs of variable and mediator were generated, and 344 pairs of indirect associations. The size of the test set for S05 is 31,520 sentences. The precision is 0.69, recall is 0.5 - slightly higher than those for S00, but still low. The overall evaluation for S05 is F1=0.59. Better results for S05 are obtained because it contains ORs written by specialist in Endocrinology, thus the text is more consistent.

6 Conclusion and Further Work

This paper reports work in progress for patient status extraction from clinical text. The experimental result show that the proposed IARM based method can

be successfully used for this task. All generated indirect association rules contain attribute-value pairs of anatomical organs/systems and their status. Although relatively small number of generated IAR, and low recall, the method can be combined with direct association rules to improve results. IARM is data-driven and unsupervised method, i.e. when larger datasets are used for IAR generation the overall evaluation will be improved. The proposed method finds relations beyond simple terms only but also helps to identify attribute-value relations in patient status description.

For more detailed further analyses of the generated IAR can be used "human-in-the-loop" [15] approach. Patients phenotype will help to identify some specific status descriptions. This can help to improve the precision. In subclustering task also can be used "human-in-the-loop" approach to reduce the complexity and dimensionality of the search space. Some structured information concerning age, gender and demographic information of the patients can be used in the filtering process to determine different IAR depending on the patient phenotype. Another direction for further work is to investigate different measures for polarity between pairs of items.

In the terminology extraction tasks [14] there are used successfully many artificial intelligence (AI) approaches: linguistic, statistical, hybrid [19], neural networks, machine learning [7], etc. The main reason for choosing IARM method is that in healthcare data processing the most important characteristic of the used method is the result to be explainable, e.i. so called "Explainable AI" [10]. This will make the decision making process more transparent and will allow further generalization of the results.

Acknowledgments

This research is supported by the grant SpecialIZed Data MIning MethoDs Based on Semantic Attributes (IZIDA), funded by the Bulgarian National Science Fund in 2017–2019. The team acknowledges the support of Medical University - Sofia, the Bulgarian Ministry of Health and the Bulgarian National Health Insurance Fund.

References

1. Abdullah, Z., Herawan, T., Ahmad, N., Ghazali, R., Deris, M.M.: Mining indirect least association rule from students' examination datasets. In: International Conference on Computational Science and Its Applications. pp. 783–797. Springer (2014)
2. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB. vol. 1215, pp. 487–499 (1994)
3. Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Integrating data analysis tools for better treatment of diabetic patients. CEUR Workshop Proceedings **2022**, 229–236 (2017)

4. Boytcheva, S., Nikolova, I., Angelova, G.: Mining association rules from clinical narratives. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 130–138 (2017)
5. Boytcheva, S., Nikolova, I., Angelova, G., Angelov, Z.: Identification of risk factors in clinical texts through association rules. In: Proceedings of the Biomedical NLP Workshop associated with RANLP. pp. 64–72 (2017)
6. Chen, L., Bhowmick, S.S., Li, J.: Mining temporal indirect associations. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 425–434. Springer (2006)
7. Conrado, M., Pardo, T., Rezende, S.: A machine learning approach to automatic term extraction using a rich feature set. In: Proceedings of the 2013 NAACL HLT Student Research Workshop. pp. 16–23 (2013)
8. Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.W., Tseng, V.S.: Spmf: a java open-source pattern mining library. The Journal of Machine Learning Research **15**(1), 3389–3393 (2014)
9. Grahne, G., Zhu, J.: High performance mining of maximal frequent itemsets. In: 6th International Workshop on High Performance Data Mining. vol. 16, p. 34 (2003)
10. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web (2017)
11. Ha, H.Y., Chen, S.C., Shyu, M.L.: Utilizing indirect associations in multimedia semantic retrieval. In: Multimedia Big Data (BigMM), 2015 IEEE International Conference on. pp. 72–79. IEEE (2015)
12. Hamano, S., Sato, M.: Mining indirect association rules. In: Industrial Conference on Data Mining. pp. 106–116. Springer (2004)
13. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data mining and knowledge discovery **8**(1), 53–87 (2004)
14. Heylen, K., De Hertog, D.: Automatic term extraction. Handbook of Terminology **1**(01) (2015)
15. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics **3**(2), 119–131 (2016)
16. Kang, S.M., Wagacha, P.W.: Extracting diagnosis patterns in electronic medical records using association rule mining. International Journal of Computer Applications **108**(15) (2014)
17. Kazienko, P.: Mining indirect association rules for web recommendation. International Journal of Applied Mathematics and Computer Science **19**(1), 165–186 (2009)
18. Manimaran, J., Velmurugan, T.: A survey of association rule mining in text applications. In: Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on. pp. 1–5. IEEE (2013)
19. Nakagawa, H., Mori, T.: A simple but powerful automatic term extraction method. In: COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14. pp. 1–7. Association for Computational Linguistics (2002)
20. Nakov, P.: Bulstem: Design and evaluation of inflectional stemmer for bulgarian. In: Workshop on Balkan Language Resources and Tools (Balkan Conference in Informatics) (2003)
21. Névéol, A., Dalianis, H., Velupillai, S., Savova, G., Zweigenbaum, P.: Clinical natural language processing in languages other than english: opportunities and challenges. Journal of biomedical semantics **9**(1), 12 (2018)

22. Tan, P.N., Kumar, V.: Mining indirect associations in web data. In: International Workshop on Mining Web Log Data Across All Customers Touch Points. pp. 145–166. Springer (2001)
23. Tan, P.N., Kumar, V.: Discovery of indirect associations from web usage data. In: Web Intelligence. pp. 128–152. Springer (2003)
24. Tan, P.N., Kumar, V., Srivastava, J.: Indirect association: Mining higher order dependencies in data. In: European Conference on Principles of Data Mining and Knowledge Discovery. pp. 632–637. Springer (2000)
25. Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., Ananiadou, S.: Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics **27**(13), i111–i119 (2011)
26. Wan, Q., An, A.: An efficient approach to mining indirect associations. Journal of Intelligent Information Systems **27**(2), 135–158 (2006)
27. Wright, A., Chen, E.S., Maloney, F.L.: An automated technique for identifying associations between medications, laboratory results and problems. Journal of Biomedical Informatics **43**(6), 891 – 901 (2010). https://doi.org/10.1016/j.jbi.2010.09.009