mathematics

MDPI

Article

WINFRA: A Web-Based Platform for Semantic Data Retrieval and Data Analytics

Addi Ait-Mlouk, Xuan-Son Vu and Lili Jiang *

Department of Computing Science, Umeå University, 90187 Umeå, Sweden; addi.ait-mlouk@cs.umu.se (A.A.-M.); sonvx@cs.umu.se (X.-S.V.)

* Correspondence: lili.jiang@cs.umu.se

Received: 21 September 2020; Accepted: 18 November 2020; Published: 23 November 2020

Abstract: Given the huge amount of heterogeneous data stored in different locations, it needs to be federated and semantically interconnected for further use. This paper introduces WINFRA, a comprehensive open-access platform for semantic web data and advanced analytics based on natural language processing (NLP) and data mining techniques (e.g., association rules, clustering, classification based on associations). The system is designed to facilitate federated data analysis, knowledge discovery, information retrieval, and new techniques to deal with semantic web and knowledge graph representation. The processing step integrates data from multiple sources virtually by creating virtual databases. Afterwards, the developed RDF Generator is built to generate RDF files for different data sources, together with SPARQL queries, to support semantic data search and knowledge graph representation. Furthermore, some application cases are provided to demonstrate how it facilitates advanced data analytics over semantic data and showcase our proposed approach toward semantic association rules.

Keywords: heterogeneous data federation; RDF; knowledge graph; data mining; natural language processing; association rules

1. Introduction

Semantic Web is a technology that aims to make knowledge understandable and machine-readable on the Web. Data in the semantic Web is structured in the formats of triple called Resource Description Framework (RDF). Recently, the massive volumes of heterogeneous data need to be federated and semantically interconnected for further use such as advanced analytics and knowledge extraction. Semantic Web techniques such as RDF, SPARQL (Protocol and RDF Query Language) have been widely used. However, the primary issue is incompleteness and insufficient integrated solution. To deal with this issue, we applied data federation, semantic Web, NLP, and data mining techniques to develop a federated data system and proposed a new approach for semantic association rules extraction. The system allows users to interact with different data sources through SPARQL queries to do advanced data analytics and interactively visualize the result.

This paper extends the work presented in [1] in the following aspects. Firstly, an up-to-date literature review was added, such as classification based association rules. Secondly, mining semantic transaction and semantic association rules by using NLP (e.g., Named Entity Recognition). The extracted association rules have also been compared to rules extracted by Apriori algorithm to highlight the value of the proposed approach and the necessity of considering text data for graph completion and new facts generation.

In this context, we processed myPersonality dataset [2] , the largest research databases in social science, collected from over 6 million volunteers on Facebook (FB). Hence, we used four of its data sources, including demographic dataset, personality dataset, political views, FB status updates dataset, and community detection dataset. Our key contributions are:

- Build up a federation system, mapping the multiple heterogeneous, distributed, and autonomous data sources into a unified federated database system, where user can choose data sources in their area of interest.
- Provide data analytics, including data exploration, which empowers users to explore the data via data mining algorithms (e.g., association rules, classification, clustering, semantic association rules), and search by queries to lead advanced analytics.
- Propose an approach to extract semantic association rules based on named entity recognition.
- Implement interactive visualization, which allows users to plot the result intuitively.
- Scale the system by adding other data sources, applying other data mining algorithms, and aiming at other data analytic scenarios.

The remainder of this paper is organized as follows. Section 2 presents a survey of related works. Section 3 explains the architecture and describes all features of WINFRA and highlight our techniques for RDF Generator, and semantic association rules. In Section 4, we describe myPersonality dataset. In Section 5, we demonstrate the results obtained through real use cases on myPersonality data. Finally, Section 6 concludes the work and outlines future work.

2. Related Work

Early researches on semantic data are based on inductive logic programming (ILP) [3] to learn new patterns. However, most of those research are not able to identify hidden patterns and relationships that data mining algorithms would. Since the new social network data are distributed in many sources, exploiting this data is a challenge, and the traditional data search approaches may produce unsatisfactory results (ignoring semantic relations). The semantic Web allows data to be used, readable by machines, and shared across applications. It empowers new capabilities to understand and retrieve new knowledge using SPARQL query language to access the data stored in the RDF graph. Querying semantic data is an important task; hence, many existing solutions provide a user-friendly interface for browsing data and allow users to perform some tasks on it. Several of these solutions are described as follows.

In the context of data federation, many enterprise tools have been developed, among them IBM InfoSphere Federation Server (ibm.co/2qWQbom) which presents enterprise data to end-users as if they were accessing a single source. Another tool is Oracle Data Service Integrator (https://goo.gl/6MKXkF) which provides a design approach to defining data transformation and integration processes. Besides, some open-source frameworks have been proposed, among them Teiid (teiid.jboss.org) which is a real-time integration engine that allows applications to use data from multiple, heterogeneous data stores. Similar efforts in data federation have also been seen from academia, such as BioMart (ensembl.org/biomart) which enables retrieval of large amounts of data in a uniform way without the need to know the database schemas; and Maelstrom which offers guidance to document and disseminate study metadata across collaborating institutions [4]. Regarding Linked Data (LD), various works have been proposed in the literature [5–7]. Besides, several tools offering RDF and linked data visualization have been developed, e.g., Sgvizler [8], LODWheel [9], IsaViz [10]. RDF-Gravity [11], RML [12], etc. Furthermore, many researchers used federated SPARQL queries to analyze and visualize linked open data [13]. However, considering advanced data analytics across federated data is ignored. Many researchers recently combined the Semantic Web (SW) and data mining techniques to improve RDF data and knowledge graph representation. Most of them on mining SW are focused on Inductive

Logic Programmings (ILP) such as WARMER [14] and ALEPH [15]; these approaches are based on ILP to generate association rules. Galarraga et al. [16,17] proposed an approach called AMIE and AMIE+ to generate closed association rules from RDF. Moreover, Molood et al. proposed a new approach called SWARM [18], which is based on AMIE and considers both the knowledge from schema level and instance level to enrich and classify extracted semantic association rules.

Ibukun et al. [19] demonstrate a procedure for improving the performance of ARM in text mining by using domain ontology. This approach reports a procedure for extracting association rules from text. However, it is based on domain ontology and keyword extraction (co-occurrences) and can not capture the semantic meaning for text. Another rule mining approach over RDF data [20] was proposed to discover association rules in RDF-based medical data. It takes the advantage of the schema-level (i.e., Tbox) knowledge encoded in the ontology to derive appropriate transactions that will later feed traditional association rules algorithms. Marinica et al. [21] proposed an interactive framework, called ARIPSO (Association Rule Interactive post-Processing using Schemas and Ontologies). The framework assists the user throughout the analyzing task to prune and filter discovered rules by using a Domain Ontology over a database. Another approach that used LOD has been proposed by Huang et al. [22] to interpret the results of text mining. The approach starts with extracting entities and semantic relations from text documents then, find frequent patterns by applying a sub-graph discovery algorithm. Another approach that uses ontologies in rule mining is the 4ft-Miner tool [23]. The tool is used in four stages of the KDD process: to map ontologies and process them to fit the standard task of association rules. All these approaches are based on ontologies which ignore semantic and dependency in text data.

In this paper, our proposed approach integrates NLP techniques and data mining (i.e., association rules mining) to extract new entity relations and semantic association rules from linked data in a federated way. It can be used for general or specific purposes (e.g., semantic information retrieval, knowledge graph completion, etc.).

3. Proposed Approach

The proposed framework is shown in Figure 1 with three major processes, including *data federation*, *data linkage*, and *knowledge discovery*. On the server-side, data from multiple data sources were preprocessed and connected through a VDB in Teiid server (teiid.io). Different techniques are firstly applied to process raw data, including generating RDF (RDF Generator) from raw data and indexing text-based data. Association rule analysis is applied to support knowledge discovery, such as exploring hidden patterns and co-occurrences of variables from multiple data sources in intuitive ways with visualization. After data exploration, users can explore the user text data in more detail and find relations between users based on their written texts by using RDF and NLP. Afterwards, users can go to the module of Data Search and issue queries across RDF endpoints for general/specific data analytics to confirm the hidden patterns they found at a large scale on all user records (i.e., 1.4M records). On the end-user side, users can explore metadata of all data sources, and further interact with them through SPARQL queries to get data analysis results.

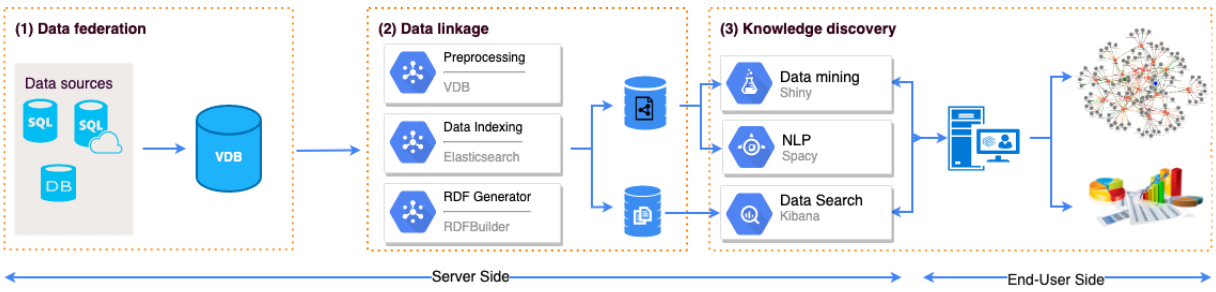

Figure 1. Overview of system architecture: (1) *Data Federation*, (2) *Data Linkage*, and (3) *Knowledge Discovery*.

The proposed approach was implemented using R, a language and environment for statistical computing and graphics. The data federation mechanism was built on Teiid. We developed the interactive and user-friendly interfaces using Shiny (shiny.rstudio.com/) and ArulesViz [24], Figure 2 presents the user interface of the system. For RDF storage, we built an RDF Generator to generate RDF files from VDB (Virtual Data Bases) and stored them in Apache Jenna Fuseki (jena.apache.org). Furthermore, to ensure high-performance data retrieval, we used ElasticSearch and Kibana.

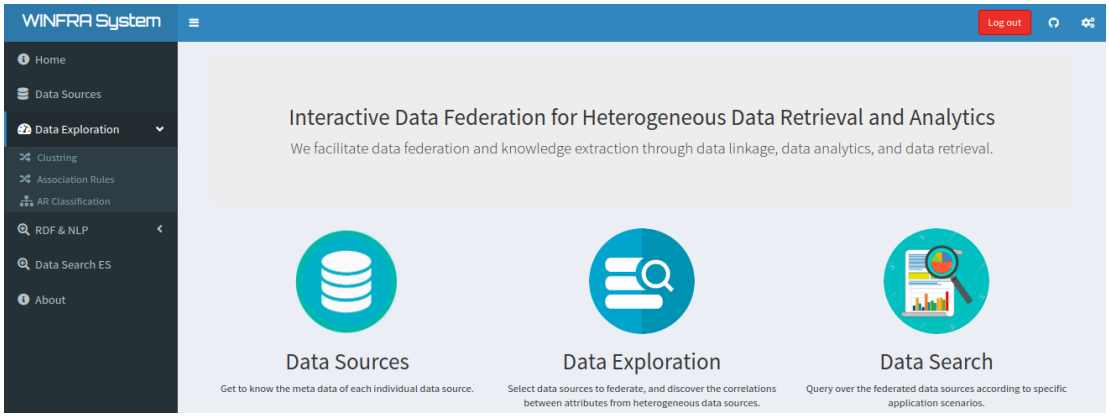

Figure 2. User interface of the WINFRA framework.

Data Federation. In this step, we created a VDB from different sources by using Teiid framework, which is a data virtualization system that allows applications to use data from multiple, heterogeneous data stores. The data is accessed and virtually integrated in real-time across distributed data sources without copying or moving data from its original location.

Data Linkage. It is a method of linking information from different sources into a single population that enables the construction of a chronological sequence of information. In this context, we applied semantic Web technologies (e.g., RDF, SPARQL) for data linkage, and our RDF Generator converts raw data to a unified RDF format as shown in Figure 1. Since we have multiple data sources federated in different locations and formats, the RDF Generator automatically maps required variables based on user queries, to return related records for further data analysis. Afterwards, we applied the inverted indexing schema from ElasticSearch (www.elastic.co) (ES) for data indexing.

Knowledge Discovery facilitates data exploration for users; we applied data mining and NLP techniques (i.e., named entity recognition) to extract semantic association rules. The following two sections will respectively explain associations rule extraction and semantic association rule mining in detail.

3.1. Association Rules Extraction and Classification

Association rules extraction. After understanding the data sources, users are enabled to use data mining algorithms to extract patterns and correlations over data variables. We took the association rules [25] technique as an example to show how data mining techniques discover the relationship between variables in federated data. The Apriori algorithm [25] was applied to extract association rules among variables over the federated data and to generate association rules. For example, the variables age (e.g., 31–40), gender (e.g., female), and relation status (e.g., married), can present an association rule graph with another variable "personality"(e.g., neuroticism, low-score agreeableness, etc.).

Classification based on association rules (CBA) [26]. Classification is one of the main techniques of data mining and machine learning. The idea is to utilize frequent patterns and relationships between objects and class labels in training data set to build a classifier. For classification based association rules in

myPersonality dataset, there is a pre-determined target (i.e., the class such as democrat, republican, etc.). The rule classification is done by focusing on the right-hand-side (RHS) subset as a restricted class attribute; we refer to this subset of rules as the class association rules. Let D be the dataset, and $I = \{i_1, i_2, \ldots, i_n\}$ a set of items, and Y be the set of class labels. A class association rule set ($CARs$) is a subset of association rules with classes specified as their consequences. The rule $X \rightarrow y$ has support value of s in D if s of the cases in D contain X and are labeled with class y. The CBA algorithm consists of two parts, a rule generator, which is based on the Apriori algorithm [25], and a classifier builder (called CBA-CB) using $CARs$ generated by rule generator. To build a classifier, let $R = \{CARs\}$ and D a training dataset, the idea is to choose a set of high precedence rules (r_i has a higher precedence than r_j if $conf(r_i) > conf(r_j)$ or $conf(r_i) = conf(r_j)$ *and* $sup(r_i) > sup(r_j)$) in R to cover D (see the illustration section for case study).

3.2. Mining Semantic Association Rules

Linked data is mostly presented in the RDF triples (SPO). However, these data are incomplete in reality; it requires promising techniques to explore and extract new facts, especially from text data. Association rule mining is a promising approach to generate such new relations, as we show in this paper. We proposed an approach (Figure 3) based on association rules mining by using NLP techniques to generate new relations from text data that can be used to enrich knowledge bases and knowledge graph representation, particularly myPersonality knowledge base.

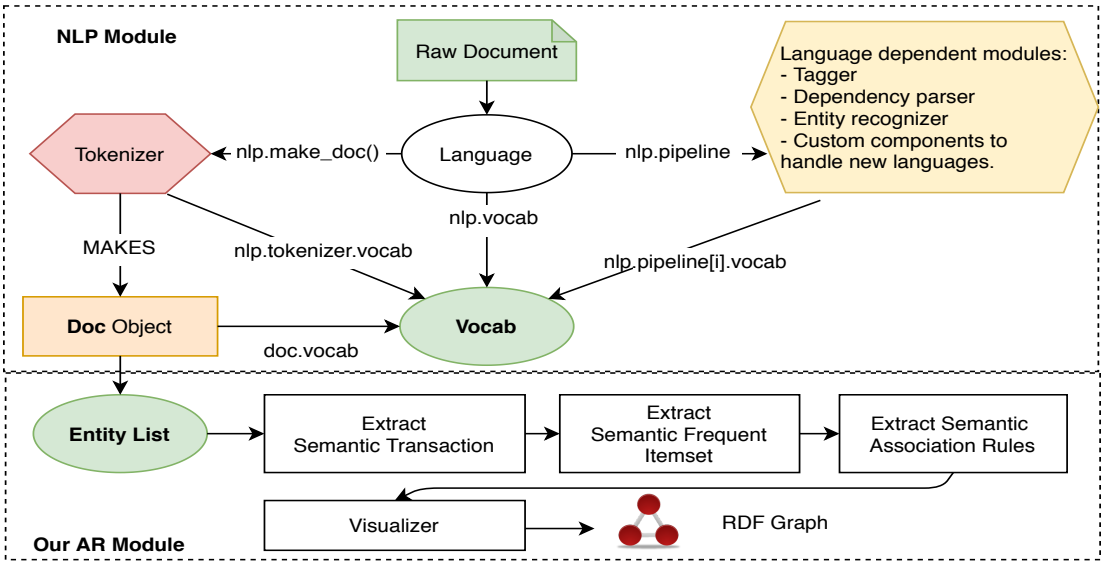

Figure 3. Mining semantic association rules from text data using Named Entity Recognition (NER) with the NLP Module built on top of SpaCy's architecture. The flexible architecture of SpaCy allows WINFRA to support multiple languages including the ability to add custom components such as custom word embeddings, to handle new terms on social media.

As shown in Figure 3, the given data (i.e., FB Status Updates—FB Posts) is processed by the NLP pipeline empowered by SpaCy NER module (spacy.io), hereafter, the NLP module. The reason we chose SpaCy are: (1) it supports multiple languages; and (2) it allows us to customize and train the NLP models to adapt new datasets. Moreover, two peer-reviewed papers [27,28] confirmed that SpaCy offers the fastest syntactic parser and reasonable accuracy, as shown in Table 1. Besides, Table 2 presents a comparison between different NLP libraries in terms of features.

Table 1. Accuracy comparison between Spacy and CoreNLP.

System	Accuracy	Speed (ms)
Spacy	91.8	13.96
CoreNLP	89.6	8.60

Table 2. Feature comparison between different NLP libraries.

Feature	Spacy	NLTK	CoreNLP
Neural network models	✓	×	✓
Integrated word vectors	✓	×	×
Multi-language support	✓	✓	✓
Tokenization	✓	✓	✓
Part-of-speech tagging	✓	✓	✓
Sentence segmentation	✓	✓	✓
Dependency parsing	✓	×	✓
Entity recognition	✓	✓	✓
Entity linking	✓	×	×
Coreference resolution	×	×	✓

More about the NLP module in Figure 3, starting from the Language object, it coordinates other components, including the Tokenizer, an NLP pipeline (Tagger, TextCategorizer, Entity Recognizer, etc.). This module takes the central data which is a text corpus (raw data) and returns an annotated document. The Doc object owns the sequence of tokens and all their annotations, while the Vocab object owns a set of look-up tables that make common information available across documents. By centralizing strings and lexical attributes, SpaCy avoids storing multiple copies of this data. This is another reason why we chose to use SpaCy to save the memory of the whole system. The final list of extracted entities and their types are stored in the Doc object. After the NLP Module, the Entity List was achieved based on multiple FB Posts. Then, it is sent to the AR module, which includes (1) Semantic Transaction Extractor, (2) Semantic Frequent Itemset Extractor, (3) Semantic Association Rules Extractor, then finally, (4) the Visualizer for displaying the RDF graph. The extraction of semantic association rules from the FB Posts is important for different research topics such as personality prediction, emotional and sentiment analysis etc [29].

To the former reason, the NLP Module allows our system to support other languages to apply our proposed approach. The myPersonality dataset contains FB Posts of users in many different countries [30]. And up to date, SpaCy has supported ten languages, such as English, German, French, etc. Therefore, based on SpaCy, our framework can process and understand multiple languages existing in user text data. To the latter reason, from the fact that user texts usually are very noisy; therefore, the ability to customize the NLP Module is a key factor for us to improve the system in the future. However, to support SpaCy in the R framework, we had to solve some software engineering problems, including how to integrate and run python inside R seamlessly. Finally, we showed that this challenge is possible to solve, which makes our system more robust to adapt to new requirements in the future.

Improve KG by using extracted entities. This approach provides an effective way to describe entities and their relationships. We use extracted entities from texts since they contain huge information from a variety of sources that can be used in semantic search, question answering, pattern mining, and sentiment analysis. In this work, we focus on the myPersonality dataset, especially FB Posts, using NER and data mining algorithms. We apply these techniques for constructing KG, including personality knowledge extraction, sentiment analysis, and community detection. Moreover, We link the extracted entities from FB posts to some external knowledge bases (i.e., DBpedia [31] and Google knowledge graph). Table 3 presents

a non-exhaustive list of extracted entities from text updates to be used in the next step for extracting semantic association rules.

Table 3. Extracted entities from FB Posts.

User ID	Entity	Entity Type
userid_0001	Nicole Gallagher	PERSON
userid_0002	Zoey	PERSON
userid_0002	Jesus	PERSON
userid_0002	Sweet	ORG
userid_0002	America	GPE
userid_0003	Karen	PERSON
userid_0004	Canadian	NORP
userid_0004	Spanish	LANGUAGE
userid_0004	World War I	EVENT
userid_0005	the Hudson River	LOC

Mining semantic association rules from entities. Users are guided to enhance the analysis by using data mining techniques to explore patterns and correlations between extracted entities. We take the association rules technique as an example to show how data mining techniques discover the relationship between different entities in the FB Posts data. This step lets the user grasp newly discovered hidden relationships between the user's FB posts that can be used to enrich KG representation.

To extract semantic association rules, semantic transaction $S_t = \{s_1, s_2, \ldots, s_n\}$ can be generated from the Entity List by using the proposed algorithm 1. The algorithm takes all FB Posts as the input and then generates entities (Table 3) and semantic transaction list (STL) as the output. It extracts a list of entities E for each FB post and stores them in E_s (lines 4–7), for each FB post, we extract a semantic transaction (ST) by generating a new subject and a list of its entities (lines 8–12). When there are no more entities, the algorithm creates a new subject and adds it to the semantic transaction list. The algorithm finally returns the list of ST (Subject, Object). After generating a semantic transaction, we fit the dataset to traditional algorithms such as Apriori [25] to generate frequent semantic itemset and then generate semantic association rules (lines 13–15). This step requires some thresholds, including minimum support and the minimum confidence to evaluate the extracted rules.

4. Dataset: myPersonality

myPersonality was a Facebook App created by David Stillwell in 2007 to allow users to participate in psychological research by sharing a personality questionnaire. Soon, over 6 million users completed the most popular questionnaire to donate their data to psychological research [2]. For demonstration, Figure 4 shows the structure of the myPersonality dataset we used.

Let's say Alice, a psychologist researcher, who wants to research the relation of personality and stress in people's lives based on social network behaviors. Alice is enabled to apply given data mining algorithms (i.e., association rules and CBA) to explore the hidden patterns through the selected variables of interests. Next, Alice will further extend the returned results in a data search to have more details on those people. Afterwards, Alice will process text data and extract semantic association rules interactively to infer new hidden information to complete and improve the knowledge graph about various case studies. These steps will be described in the following case studies accordingly.

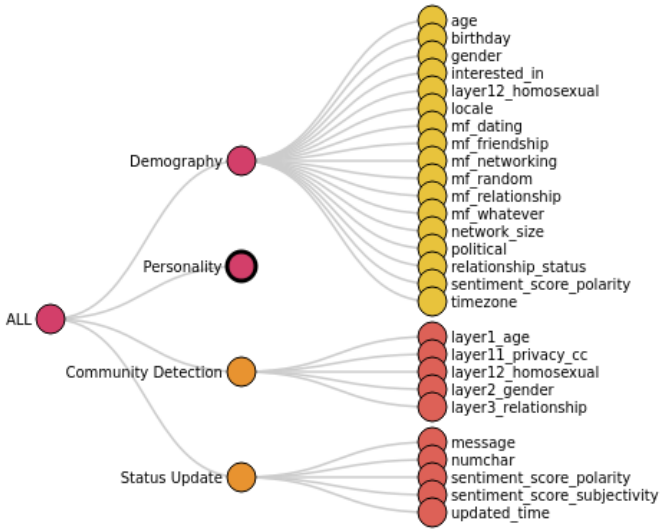

Figure 4. Structure of myPersonality dataset.

Algorithm 1: Semantic association rules (SAR) from text data

Input : FB Posts, *minisupp*, *minconf*
Output: SAR

$S \leftarrow \emptyset, O \leftarrow \emptyset, STL \leftarrow \emptyset, list \leftarrow \emptyset$
foreach *Post* $S_i \in$ *FB Posts* **do**
  $E_s \leftarrow \emptyset$
  foreach $t_j \in$ *Post* S_i **do**
    $E_j \leftarrow NER(S_j)$
    $E_s \leftarrow E_s + E_j$
  end
  $S \leftarrow Post(user)$
  $O \leftarrow E_s$
  $ST \leftarrow \bigcup(S, O)$
  $STL \leftarrow list(ST)$
end
foreach $ST_i \in STL$ **do**
  $SAR \leftarrow$ APRIORI(STL)
end
return SAR

5. Illustrative Examples

5.1. Case-Study 1: Interactive Association Rules Based on Selected Variables

In this case study, Alice will select her variables of interests including sentiment_score_subjectivity (sentiS), cNEU (neuroticism), cCON (conscientiousness), cAGR (agreeableness) [32] in the graph visualization panel ① to be analyzed through association rules. Afterwards, we applied the Apriori algorithm to extract frequent itemsets and then generate association rules based on the minimum

confidence defined by the user. As shown in Figure 5, in data exploration, the four variables were chosen to extract association rules. On the right side, the configuration panel ② was displayed to control the mining process by setting data source, support(*minsupp*), and confidence(*minconf*) thresholds. For instance, we fixed the value of *minsupp* = 0.4 to generate all possible frequent itemsets and *minconf* = 0.5 to generate and filter interesting rules from frequent itemsets previously extracted. Afterwards, Alice moves to the bottom panel ③ and clicks on the "association rule graph" tab to see the results like in Figure 6. In this Figure, the rectangles represent variables, and the circles represent association rules. The larger size of the circle implies more data records matching the rule, while the darker circle represents more importance of the rule.

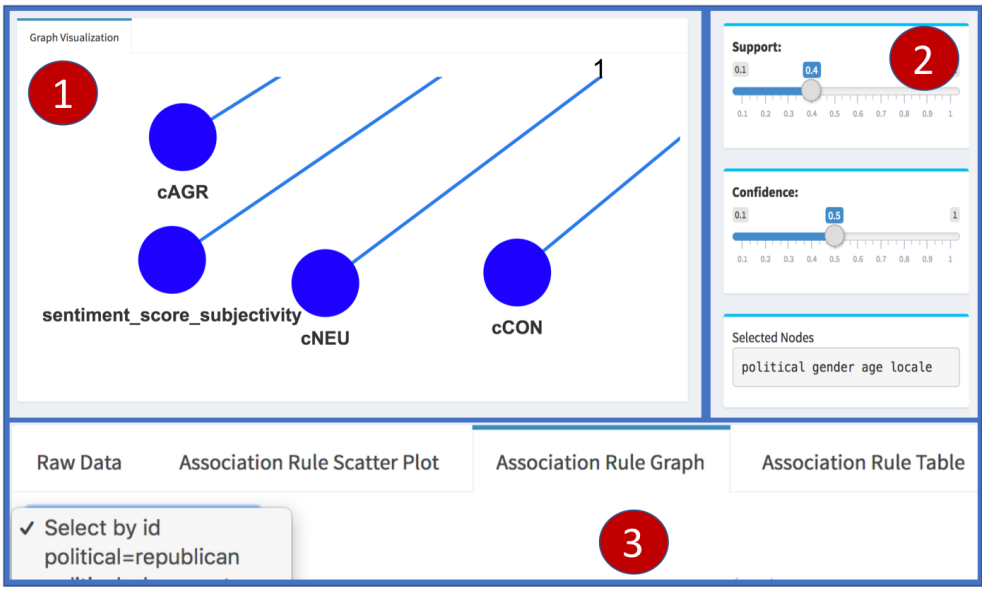

Figure 5. Selected variables on *Data Exploration*.

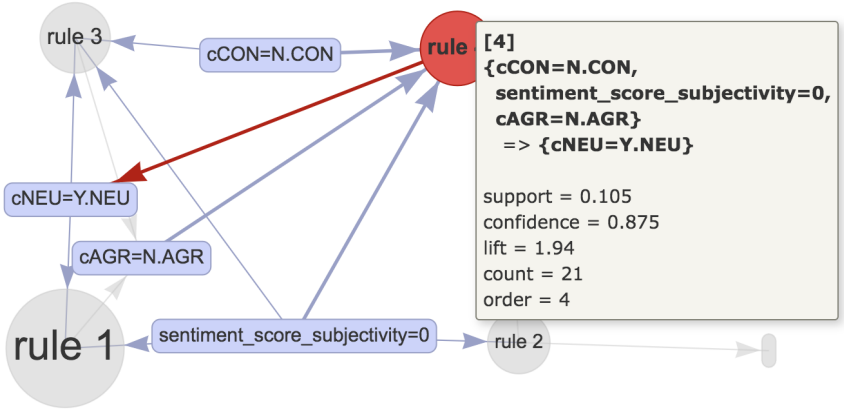

Figure 6. Case study of association rule based on selected variables.

Regarding the research question, Alice found a hidden pattern between three variables regarding neurotic people (cNEU), which is a personality trait that reflects one's ability to deal with emotional states, such as stress and anxiety. The pattern suggests that people who are not agreeableness (N.AGR) and do not have strong purposes in the way they say on FB status (sentiS = 0) are most likely neurotic people.

5.2. Case-Study 2: Classification Based on Association Rules (CBA)

The purpose of CBA is to classify rules that satisfy minimum support and minimum confidence thresholds. It generates all candidate *k*-itemsets and then calculates the support for finding frequent itemsets that satisfied *minsupp*. Then it generates all candidate *k*-itemsets from the frequent (*k*-1)-itemsets in a similar way to the Apriori algorithm. The rule is relevant if the confidence is greater than *minconf*. The set of class association rules (*CARs*), therefore, consists of all the possible rules that are both frequent and relevant. In this case study, firstly, we use political view dataset to analyze and extract association rules between variables related to myPersonality dataset. Next, the extracted rules are used to build a classifier based on relevant rules (higher confidence). For instance, users can interact directly with the system to change thresholds and control more the quantity of extracted rules. Classification based rule task is an interesting practical application. For example, when classifying rules about the political view (democrat, centrist, independent, etc.), including demographic information such as gender and age in the rule antecedent. For instance the rule with antecedent:

$(cCON = N.CON \wedge cNEU = Y.NEU \wedge locale = en\_US \wedge relationship\_status = 3)$ classified users to republican political view. Some rules with very high confidence and lift but with low support.

For example in Table 4, and according to specified thresholds for support and confidence, the rules *{cOPN = Y.OPE, gender = 1, layer2_gender = 0, timezone = −4}* and *{political = democrat}* suggest that females who are openness and not connected to other people are classified to democrat.

Table 4. Example of rules based classification on myPersonality (political view dataset).

N	LHS	RHS
1	$\{cOPN = Y.OPE, cAGR = N.AGR, timezone = -5, layer3\_relationship = 1\}$	$\{political = democrat\}$
2	$\{cCON = Y.CON, cOPN = Y.OPE, gender = 1, location = en\_us, layer2\_gender = 0\}$	$\{political = democrat\}$
3	$\{layer12\_homosexual = -99, timezone = -4, entiment\_score\_subjectivity = 1\}$	$\{political = democrat\}$
4	$\{cOPN = Y.OPE, gender = 1, interested\_in = -99, timezone = -4\}$	$\{political = democrat\}$
5	$\{cOPN = Y.OPE, gender = 1, layer2\_gender = 0, timezone = -4\}$	$\{political = democrat\}$
6	$\{cNEU = N.NEU, age = 19, gender = 0, mf\_friendship = -99\}$	$\{political = doesn't care\}$
7	$\{sCON = 2.75, sentiment\_score\_polarity = 0\}$	$\{political = democrat\}$
8	$\{sOPN = 4.25, relationship\_status = 1\}$	$\{political = doesn't care\}$

5.3. Case-Study 3: Mining Semantic Association Rules

In this case study, we show how entities were extracted from the FB Posts and how semantic rules were extracted. We applied the proposed algorithm to prepare semantic transactions through entity relation extraction. Next, we used the Apriori algorithm on the generated semantic transaction to extract frequent semantic itemset, which satisfies the minimum support requirements defined by the user and then generate semantic association rules based on the user-defined confidence threshold. As shown in Table 5, the result does not present useful information to the users since the data is text (status update), and this is a major issue for basic association rules mining. The proposed approach tackled this issue by extracting new semantic association rules and compared it with the early work [1], the result presented in Table 5 and Figure 7. The result of the proposed approach (Figure 8, Table 6) shows the ability to extract new semantic rules from FB Posts and new relationships between different objects and subjects (OS). The new extracted relations can be used to improve RDF data and generate new facts for knowledge base completion and knowledge graph (KG) representation.

Table 5. Top 8 rules extracted from FB posts.

N	LHS	RHS	S	C	L
1	$\{userid = userid\_0001\}$	{status= Another day of life.. how boring. Summer's making me leak[...]}	0.01	1	100
2	$\{userid = userid\_0002\}$	{status= Watchin tv. reli bored this weekend is going way to fast [...]}	0.01	1	100
3	$\{userid = userid\_0003\}$	{status= Is so in love...want boyfriend here...and we can[...]}	0.01	1	100
4	$\{userid = userid\_0004\}$	{Most people have 1000 wishes for Christmas; a cancer patient only [...]}	0.01	1	100
5	$\{userid = userid\_0005\}$	{Just back from charleston, that edifi meeting was a bust! hankful [...]}	0.02	1	50
6	$\{userid = userid\_0006\}$	{status=christmas party daw ng MN4-4 sa december 05, hopefully[...]}	0.02	1	50
7	$\{userid = userid\_0007\}$	{status=The injustice of college textbooks: 105 for a lab manual[...]}	0.02	1	50
8	$\{userid = userid\_0008\}$	{status=Semangat buat yang mau ujian. Pertempuran besar akan segera[...]}	0.02	1	50

Table 6. Top extracted semantic association rules from FB posts using NER approach.

N	LHS	RHS	S	C	L	CF
1	$\{userid = userid\_0008\}$	$\{entity\_type = PERSON\}$	0.06	0.6	2.1	0.4
2	$\{userid = userid\_0004\}$	$\{entity\_type = PERSON\}$	0.06	0.6	2.1	0.4
3	$\{userid = userid\_0010, entity = Karen\}$	$\{userid = userid\_0010\}$	0.10	1.0	74	1.0
4	$\{entity\_type = GPE, entity = Karen\}$	$\{entity\_type = PERSON\}$	0.06	1.0	2.1	0.4
5	$\{entity\_type = PERSON, entity = Erin\}$	$\{userid = userid\_0008\}$	0.08	1.0	10	1.0
6	$\{entity\_type = GPE, entity = Cya\}$	$\{userid = userid\_0009\}$	0.04	1.0	4.3	1.0
7	$\{userid = userid\_0009, entity = Cya\}$	$\{entity\_type = GPE\}$	0.04	0.9	4.8	0.9
8	$\{userid = userid\_0009, entity\_type = GPE\}$	$\{entity = Cya\}$	0.04	0.7	16	0.7
9	$\{entity\_type = GPE, entity = Cya\}$	$\{userid = userid\_0009\}$	0.04	1.0	4.3	1.0
10	$\{userid = userid\_0009, entity = Cya\}$	$\{userid = userid\_0009\}$	0.04	1.0	4.3	1.0
11	$\{entity\_type = GPE, userid = userid\_0010\}$	$\{entity = Cya\}$	0.04	1.0	4.3	1.0

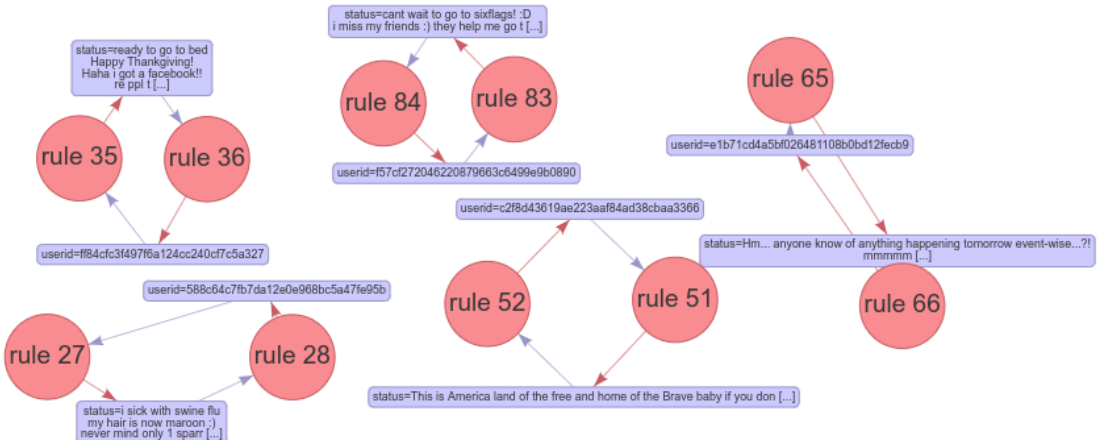

Figure 7. Some association rules before using proposed approach.

5.4. Evaluations

Various quality measurements have been used to evaluate and select the most significant association rules, such as Support (S), Confidence (C) and Lift (L). Association rules are about finding patterns in data, it is not a classification problem, and the accuracy measure does not make sense and does not provide any evaluation. However, it is possible to use a certainty factor (CF) [33] to assess and evaluate the extracted association rules. It is an alternative of accuracy measure for association rules.

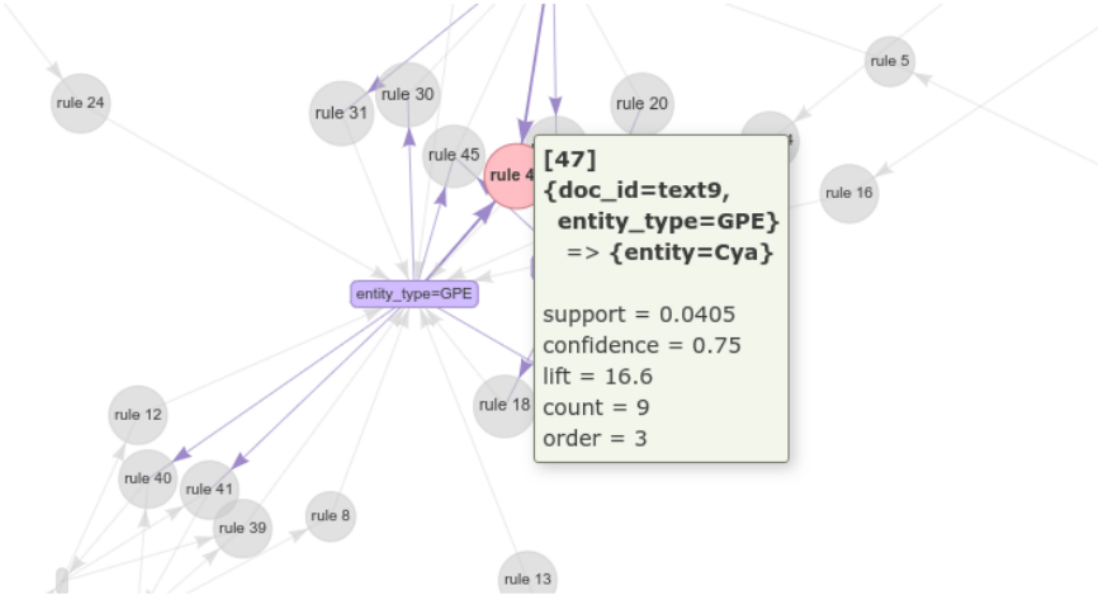

Figure 8. Some semantic association rules after using the proposed approach.

Certainty factors are used to represent uncertainty in the rules in expert systems, and it measures the variation of the probability that B is in a transaction when only considering transactions with A. An increase of CF means a decrease in the probability that B is not in a transaction that A is in. The CF of a given association rule ($A \rightarrow B$) is based on its confidence and support; the formula is given in Equation (1).

$$CF(A \rightarrow B) = \begin{cases} \frac{C(A \rightarrow B) - S(B)}{1 - S(B)} \; if \; C(A \rightarrow B) > S(B) \\ \frac{C(A \rightarrow B)}{S(B)} \; otherwise \end{cases} \tag{1}$$

$S(B)$ is computed by using the confidence and lift of rule using Equation (2):

$$S(B) = \frac{C(A \cup B)}{L(A \cup B)} \tag{2}$$

We consider an association rule $A \rightarrow B$ as strong when its support and CF are greater than thresholds *minsupp and minCF*, respectively. The evaluation of top extracted rules according to the preferences of decision-makers (e.g., *minsupp* = 0.4, *minconf* = 0.5, *minCF* = 0.3) are given in Table 6. These top extracted rules are sensitive to predefined thresholds by the user. This can help the decision-makers to evaluate and extract only the most interesting rules according to their preferences and objectives. This task of association rules can be stressful and time-consuming for the user in each update of thresholds; however, the proposed platform make it easy via interactive web interfaces. The user can update thresholds interactively to conduct their tasks (data exploration, association rules, clustering, classification, NER, text mining, etc.), as well as federated data analysis and information retrieval(IR) on linked data.

In summary, the proposed approach contributes to a better understanding of federated databases and text data (myPersonality dataset) by extracting new association rules through Apriori-like algorithms and named entity recognition (NER). The developed platform has the following three major strengths: (1) the overall WINFRA system provides advanced analysis on a federation system through SPARQL queries; (2) The newly extracted semantic association rules improve knowledge graph representation by providing new facts (RDF triples) for graph completion; (3) the overall WINFRA provides an open-access

to serve social science on myPersonality and help researchers to interact with social data in an anonymous and federated way. Besides, the task of semantic association rules is sensitive to the Named Entity Recognition task. Extracting relevant entities leads to interesting semantic transactions that can be used by the association rules pipeline to extract semantic association rules (the accuracy of semantic rules task depends on the accuracy of the NER task). For instance, the platform supports many tasks, among them data exploration, association rules, data clustering, classification based association rules, and semantic association rules from text data for knowledge graph completion as well as advanced data analysis. The platform is built on top of myPersonality Facebook dataset to answer some research questions and address particular use cases such as privacy-concern analysis, personality traits, sentiment analysis, and association rules-based sentiment analysis. This could be very useful and beneficial for social science researchers [2] who want to explore the myPersonality dataset and use SPARQL queries to query social data from WINFRA endpoint.

6. Conclusions

In this paper, we have developed WINFRA, a comprehensive open-access platform for semantic web data, and advanced analytics. It is a SPARQL query based visualizer for heterogeneous data sources (i.e., myPersonality), which allows users to conduct interactive advanced data analytics over heterogeneous data sources. Besides, to solve the issue that the existing approaches focus on triple (Subject, Predicate, Object) to generate semantic association rules (SAR) and ignore text data, we proposed a new approach to extract SAR from text data and construct semantic association rules. We applied this approach to myPersonality data (i.e., FB Posts) to process the status of different users and extract entities used by the mining step to generate new semantic association rules that can be used to improve RDF data and generate new facts for knowledge base completion and Knowledge graph (KG) representation.

In future work, we would like to propose more techniques to tackle ethical issue and privacy-guarantees of sensitive data by introducing federated learning. Furthermore, upload and index external data directly in the platform for specific use cases.

Author Contributions: Formal analysis, investigation and methodology, A.A.-M., X.-S.V., L.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Federated Database project (570066000) funded by Umeå University, Sweden.

Acknowledgments: The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at HPC2N center. The authors also thank the myPersonality project for data contribution.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The data used was applied from myPersonality project.

References

1. Vu, X.S.; Ait-Mlouk, A.; Elmroth, E.; Jiang, L. Graph-based Interactive Data Federation System for Heterogeneous Data Retrieval and Analytics. In *WWW'19, Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019*; ACM: New York, NY, USA, 2019; pp. 3595–3599. [CrossRef]
2. myPersonality, Project Home Page. Available online: https://sites.google.com/michalkosinski.com/mypersonality (accessed on 24 September 2020).
3. Muggleton, S.; de Raedt, L. Inductive Logic Programming: Theory and methods. *J. Log. Program.* **1994**, *19–20*, 629–679. [CrossRef]
4. Maelstrom, Project Home Page. Available online: https://www.maelstrom-research.org/ (accessed on 24 September 2020).

5. Brunetti, J.M.; Auer, S.; García, R. The Linked Data Visualization Model. In *ISWC-PD'12, Proceedings of the 2012th International Conference on Posters and Demonstrations Track, Boston, MA, USA, 11–15 November 2012*; CEUR-WS.org: Aachen, Germany, 2012; Volume 914, pp. 5–8.
6. Alonen, M.; Kauppinen, T.; Suominen, O.; Hyvönen, E. Exploring the Linked University Data with Visualization Tools. In *The Semantic Web: ESWC 2013 Satellite Events, Montpellier, France, 26–30 May 2013*; Cimiano, P., Fernández, M., Lopez, V., Schlobach, S., Völker, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 204–208.
7. Martínez-Rodríguez, J.L.; Hogan, A.; López-Arévalo, I. Information extraction meets the Semantic Web: A survey. *Semant. Web* **2020**, *11*, 255–335.
8. Skjæveland, M.G. Sgvizler: A JavaScript Wrapper for Easy Visualization of SPARQL Result Sets. In *The Semantic Web: ESWC 2012 Satellite Events, Crete, Greece, 27–31 May 2012*; Simperl, E., Norton, B., Mladenic, D., Della Valle, E., Fundulaki, I., Passant, A., Troncy, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; pp. 361–365.
9. Stuhr, M.; Roman, D.; Norheim, D. LODWheel—JavaScript-based Visualization of RDF Data. In Proceedings of the Second International Conference on Consuming Linked Data, COLD'11, Bonn, Germany, 23 October 2011; pp. 73–84.
10. IsaViz, Project Home Page. Available online: https://www.w3.org/2001/11/IsaViz/ (accessed on 24 September 2020).
11. rdf-gravity, project Home Page. Available online:https://www.salzburgresearch.at/publikation/rdf-gravity-3/ (accessed on 24 September 2020).
12. Meester, B.D.; Heyvaert, P.; Verborgh, R.; Dimou, A. *Mapping Languages: Analysis of Comparative Characteristics*; KGB@ESWC: Portorož, Slovenia, 2019.
13. Djokic-Petrovic, M.; Cvjetkovic, V.; Yang, J.; Zivanovic, M.; Wild, D.J. PIBAS FedSPARQL: A web-based platform for integration and exploration of bioinformatics datasets. *J. Biomed. Semant.* **2017**, *8*, 42.
14. Goethals, B.; Van Den Bussche, J. Relational association rules: Getting Warmer. In *Pattern Detection and Discovery*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 145–159.
15. Muggleton, S. Inverse entailment and progol. *New Gener. Comput.* **1995**, *13*, 245–286. [CrossRef]
16. Galárraga, L.A.; Teflioudi, C.; Hose, K.; Suchanek, F. AMIE: Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases. In Proceedings of the 22nd International Conference on World Wide Web, WWW'13, Rio de Janeiro, Brazil, 13–17 May 2013; ACM: New York, NY, USA, 2013; pp. 413–422. [CrossRef]
17. Galárraga, L.; Teflioudi, C.; Hose, K.; Suchanek, F.M. Fast rule mining in ontological knowledge bases with AMIE++. *VLDB J.* **2015**, *24*, 707–730. [CrossRef]
18. Barati, M.; Bai, Q.; Liu, Q. Mining semantic association rules from RDF data. *Knowl. Based Syst.* **2017**, *133*, 183–196. [CrossRef]
19. Daramola, O.; Ibukun, A.; Okuboyejo, O. Semantic association rule mining in text using domain ontology. *Int. J. Metadata Semant. Ontol.* **2017**, *12*, 28. [CrossRef]
20. Nebot, V.; Berlanga, R. Finding association rules in semantic web data. *Knowl.-Based Syst.* **2012**, *25*, 51–62. [CrossRef]
21. Marinica, C.; Guillet, F. Knowledge-Based Interactive Postmining of Association Rules Using Ontologies. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 784–797. [CrossRef]
22. Huang, Z.; Chen, H.; Yu, T.; Sheng, H.; Luo, Z.; Mao, Y. Semantic Text Mining with Linked Data. In Proceedings of the Fifth International Joint Conference on INC, IMS and IDC, Seoul, Korea, 25–27 August 2009; pp. 338–343. [CrossRef]
23. Svátek, V.; Rauch, J.; Ralbovský, M. Ontology-Enhanced Association Mining. In *Semantics, Web and Mining*; Ackermann, M., Berendt, B., Grobelnik, M., Hotho, A., Mladenič, D., Semeraro, G., Spiliopoulou, M., Stumme, G., Svátek, V., van Someren, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 163–179.
24. Hahsler, M.; Karpienko, R. Visualizing association rules in hierarchical groups. *J. Bus. Econ.* **2017**, *87*, 317–335.
25. Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases, VLDB'94, Santiago de Chile, Chile, 12–15 September 1994; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1994; pp. 487–499.
26. Hu, K.; Lu, Y.; Zhou, L.; Shi, C. Integrating Classification and Association Rule Mining: A Concept Lattice Framework. In *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*; Zhong, N., Skowron, A., Ohsuga, S., Eds.; Springer: Berlin/Heidelberg, Germany, 1999; pp. 443–447.

27. Honnibal, M.; Johnson, M. An Improved Non-monotonic Transition System for Dependency Parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 1373–1378. [CrossRef]
28. Choi, J.D.; Tetreault, J.; Stent, A. It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 26–31 July 2015; Association for Computational Linguistics: Beijing, China, 2015; pp. 387–396. [CrossRef]
29. Settanni, M.; Marengo, D. Sharing feelings online: Studying emotional well-being via automated text analysis of Facebook posts. *Front. Psychol.* **2015**, *6*, 1045. [CrossRef] [PubMed]
30. Vu, X.S.; Jiang, L. Generic Multilayer Network Data Analysis with the Fusion of Content and Structure. *arXiv* **2019**, arXiv:1905.08635.
31. DBpedia, Project Home Page. Available online: https://wiki.dbpedia.org/ (accessed on 24 September 2020).
32. Vu, X.S.; Flekova, L.; Jiang, L.; Gurevych, I. Lexical-semantic resources: Yet powerful resources for automatic personality classification. In Proceedings of the 9th Global WordNet Conference, Singapore, 8–12 January 2018; pp.173–182.
33. Sánchez, D.; Serrano, J.; Blanco, I.; Martin-Bautista, M.; Vila, M. Using association rules to mine for strong approximate dependencies. *Data Min. Knowl. Discov.* **2008**, *16*, 313–348. [CrossRef]