PAPER Special Section on Semantic Web and Linked Data

Classification of Linked Data Sources Using Semantic Scoring

Semih YUMUSAK[†a)], ***Student Member***, **Erdogan DOGDU**[††], ***and*** **Halife KODAZ**[†††], ***Nonmembers***

SUMMARY Linked data sets are created using semantic Web technologies and they are usually big and the number of such datasets is growing. The query execution is therefore costly, and knowing the content of data in such datasets should help in targeted querying. Our aim in this paper is to classify linked data sets by their knowledge content. Earlier projects such as LOD Cloud, LODStats, and SPARQLES analyze linked data sources in terms of content, availability and infrastructure. In these projects, linked data sets are classified and tagged principally using VoID vocabulary and analyzed according to their content, availability and infrastructure. Although all linked data sources listed in these projects appear to be classified or tagged, there are a limited number of studies on automated tagging and classification of newly arriving linked data sets. Here, we focus on automated classification of linked data sets using semantic scoring methods. We have collected the SPARQL endpoints of 1,328 unique linked datasets from Datahub, LOD Cloud, LODStats, SPARQLES, and SpEnD projects. We have then queried textual descriptions of resources in these data sets using their `rdfs:comment` and `rdfs:label` property values. We analyzed these texts in a similar manner with document analysis techniques by assuming every SPARQL endpoint as a separate document. In this regard, we have used WordNet semantic relations library combined with an adapted term frequency-inverted document frequency (tfidf) analysis on the words and their semantic neighbours. In WordNet database, we have extracted information about comment/label objects in linked data sources by using hypernym, hyponym, homonym, meronym, region, topic and usage semantic relations. We obtained some significant results on hypernym and topic semantic relations; we can find words that identify data sets and this can be used in automatic classification and tagging of linked data sources. By using these words, we experimented different classifiers with different scoring methods, which results in better classification accuracy results.

key words: *linked data, semantic classification, wordnet*

1. Introduction

Linked data sources are structured web of data sources, created using semantic web technologies and specifically a triple-based graph infrastructure [7]. These data sources are served in many different ways, such as plain RDF (Resource Description Framework) data files in different formats (N-Triples, Turtle, JSON, etc.), RDF data stores (Virtuoso, Apache Jena, OntoQuad, etc.), and SPARQL endpoints. Among these, SPARQL endpoints are designed for live querying of linked data sources by binding several different RDF data stores. However, a majority of the SPARQL endpoints do not contain any information about the inner content (i.e. knowledge base). For this purpose, there are repositories (e.g. Datahub, SPARQLES, LODStats, LOD Cloud), listing contextual information about SPARQL endpoints. Nevertheless, most of the endpoints are still not indexed and not categorized in these repositories. In an ongoing study [31], we developed a new SPARQL endpoint discovery engine. It is currently serving a larger and more comprehensive set of SPARQL endpoints than the previous studies [31]. However, without a proper description of the content, these endpoints could not be effectively used by data consumers. In that sense, classification of the Web of linked data creates an important guide for linked data consumers who make use of linked data sources in many different data access scenarios. For example, live SPARQL query language processing, federated querying, direct RDF access and querying, embedded RDFa indexing, and RDF indexing systems will make use of the contextual information about linked data sources provided in this study. Specifically for the SPARQL endpoints, federated query processing engines, such as SPLENDID [11], HiBISCuS [25], ANAPSID [4] are online query processing tools to retrieve simultaneous results from several SPARQL endpoints, which also in need for contextual information. Furthermore, Link traversal based query execution systems such as SQUIN [12] uses traversal querying of linked data sources. Above all, SPARQL endpoints are used as an entry point for most of the linked data sources, usually with a simple web page as a user interface for SPARQL querying. As for federated query engine optimization, it is necessary to classify SPARQL endpoints in order to eliminate irrelevant endpoints by their relevance to the query [25]. In order to identify and serve what an endpoint semantically contains, we propose a ranking and classification method for linked data sources together with a topic recommendation extension.

In this study, we have first unified all SPARQL endpoint URLs collected from SPARQLES [8], LOD Cloud [9], LODStats [5], SpEnD* projects (Sect. 3.2). Then, we have collected text data (Sect. 3.2) from our unified SPARQL endpoint list of linked data sources. Finally, we applied classification algorithms by using a tfidf scoring method enhanced by Wordnet semantics (Sect. 3.3).

The paper is organized as follows. The related work is reviewed in Sect. 2. Proposed data collection and classification methods are presented in Sect. 3. Preliminary results are presented and discussed in Sect. 4. Finally, we conclude in Sect. 5.

Manuscript received February 16, 2017.
Manuscript revised June 29, 2017.
Manuscript publicized September 15, 2017.
[†]The author is with KTO Karatay Univ., Turkey.
[††]The author is with Cankaya Univ., Turkey.
[†††]The author is with Selcuk University, Turkey.
a) E-mail: semih.yumusak@karatay.edu.tr
DOI: 10.1587/transinf.2017SWP0011

*https://github.com/semihyumusak/SPEC

2. Related Work

2.1 SPARQL Endpoint Sources

Meta data about the major linked data sources are collected and stored in CKAN[†] data set. LOD Project [9] and SPARQLES [8] projects also use CKAN to store data sets in the Datahub[††] web project. LODStats [5] project offers a statistical analysis on linked data sources collected from different sources, which are available through its web site. As an ongoing project, SpEnD [31] focuses on discovering new SPARQL endpoints from all over the Web using meta search techniques. All of these repositories contain many linked data sources as well as their SPARQL endpoints.

2.2 Categorization and Topic Identification

Topic modeling approaches on document based systems [30] offer text based document analysis for identifying the topic of a document. These approaches are applied on web pages [21]–[23], [27] as well as in the Linked Data domain [6]. Although there are studies on automatic classification of LOD Datasets [19], data set topic identification and classification are mainly done by manual selection and categorization [9]. In this respect, LOD Cloud diagram includes category tags for linked data sources which are based on the topic names that are manually entered by CKAN data publishers. LOD Cloud categories include nine topics, which are: *publications*, *life sciences*, *cross-domain*, *social networking*, *geographic*, *government*, *media*, *user-generated content*, and *linguistics* [9]. As [19] suggests, manual classification of data sets causes incorrect tagging when compared to a statistical classification which results in 81,62% accuracy. An earlier approach was developed by [10] on linked data classification which is based on features. Furthermore, [15] developed a similar approach by using Freebase as a topic discovery tool and applied this approach to LOD data sets. [19] and [15] focused mainly on LOD Cloud whereas [10] exemplifies the clustering algorithm with a generic feature matching clustering approach instead.

2.3 Wordnet Semantics

WordNet [20] is a lexical database for English words and includes semantic relations between words. The semantic relations are defined as synonymy, antonymy, hyponymy, meronymy, toponymy, and entailment. Hyponymy is also used as hypernymy inversely. Hypernymy is basically defined as a type relationship. For instance, A is defined as a hypernym of B, whenever B may be categorized as a type of A [3]. The exact opposite of hypernymy relationship is defined as hyponymy. Using Wordnet library, the possible topic for a word can also be extracted by requesting topic semantics relations of a single word. A complete analysis of every word in a document may help us to predict the topic of a document or create content related tags for that document.

3. Data Collection and Methodology

3.1 SPARQL Endpoint URLs

In a previous study [31], we have collected SPARQL endpoint URLs from different providers. Including our previous list of endpoints, we have mainly used five different data collections: SpEnD, LOD Project, SPARQLES, LODStats, and Datahub. While we were collecting endpoint URLs, we have found that Datahub is a common sharing platform for SPARQLES, LOD Project, and LODStats projects that use CKAN platform to publish metadata about data sets [31].

By using the crawling and community source related collection methods, we have collected and merged 1328 unique SPARQL endpoint URLs for analysis.

3.2 Textual Content Collection

We have queried and collected the textual descriptions in the linked data sets using SPARQL querying via Apache Jena[†††]. Specifically, we have collected *rdfs:comment* and *rdfs:label* property values (text) by executing the following two SPARQL queries on all SPARQL endpoints.

```
SELECT DISTINCT ?o WHERE ?s rdfs:comment ?o
SELECT DISTINCT ?o WHERE ?s rdfs:label ?o
```

The collected raw text data (in comments and labels) are cleaned from unrecognized characters, parsed and split into words in order to prepare data for word analysis. Collected data is stored in a local database. The analysis about the SPARQL endpoints, objects collected, and the words found are further listed in Sect. 4.

3.3 Semantics Based Frequency Scoring

Tf-Idf scoring [26] is a method that is used for calculating the relevancy of a word in a document within a set of documents. In document classification [28], the frequency scores are additionally used to refine the feature selection processes [13]. In this study, we have used the tf-idf score as our base for the SPARQL endpoint classification task. Then, we enhanced our document-endpoint library by using Wordnet hypernyms and topics. In the literature, Wordnet was used to improve the classification accuracy in document classification tasks [16]. By considering SPARQL endpoints as documents, we have used semantically enhanced tf-idf scores to classify SPARQL endpoints. We used tf-idf scoring and added hypernym/topic term count and term inverse endpoint count for every hypernym/topic derived for each term. As a consequence of this enhancement, a combination of standard tf-idf scoring and semantics scoring was created,

[†]http://ckan.org/
[††]http://datahub.io/
[†††]https://jena.apache.org/

which we named as Stfidf scoring. The classical tfidf scoring [29] is formalized as:

$$\text{tfidf(t, d, D)} = tf(t,d).idf(t,D)$$

where the terms are:

D: set of documents
d: a single document
t: a single term

Starting from this expression, we first replaced the `"t"` term with the `"s"` semantic term and replaced all terms with their semantically related terms as:

$$\text{Stfidf(s, } d_s, D_s) = tf(s,d_s).\text{idf(s,}D_s)$$

Where the terms are:

D_s: semantically converted documents
d_s: single converted document
s: a semantic term extracted from WordNet

In this context, D_s is the set of document created from the original documents by replacing all words with their WordNet hypernyms or topics. For example, the word `compilers` is converted to its hypernym `computer program`, and the word `infection` is converted to its topic `medicine`.

Furthermore, in order to measure the effect of this scoring in contrast to the classical tfidf scoring, a combined version of Stfsidf scoring is also experimented and it is calculated as follows:

$$\text{Ctfidf(t, s, } d_c, D_c) = \text{tfidf(t, d, D).Stfidf(s, } d_s, D_s)$$

D_c: Documents containing original words together with their hypernyms/topics as tags
d_c: a single document with semantic tags

Finally, the Stfidf scoring function results are normalized between 0 and 1 score for every word, semantic term, and endpoint.

3.4 Classification of Linked Data Sources

In order to understand the effect of the proposed tfidf scoring methods (Stfidf and Ctfidf), the scoring functions are used to create the feature vectors before running different classification methods. Within this context, SPARQL endpoints are taken as the Linked Data sources to be classified. Those linked data sources are used to create document vectors for every data source. Thereafter, the document vectors were used as the training set. The LOD Cloud [9] categories (mentioned in Sect. 2.2) are used as class labels (*publications*, *life sciences*, *cross-domain*, *social networking*, *geographic*, *government*, *media*, *user-generated content*, and *linguistics*).

Since the number of SPARQL endpoints are very limited, Leave-one-out cross validation technique [24] is used to calculate the most accurate classification results. The input parameters for the classification algorithms are tuned for this specific case and the results are calculated by using an incremental feature selection [17] method. Thereby, the effect of the scoring method can be experimented many times with different number of features.

We have experimented with seven different classification algorithms and compared the results. These classification algorithms are Ada Boost, Decision Tree, Linear SVM, Naive Bayes, Nearest Neighbors, Random Forest, RBF SVM.

4. Results and Analysis

Not all SPARQL endpoints we collected are accessible and have enough data. So, we filtered out the ones that are not useful for our study. Following is a summary of the filtering process and the results.

- 1.328 SPARQL endpoints are collected initially from the relevant collections.
- 676 of 1.328 SPARQL endpoints are accessible online and contain `rdfs:comment` or `rdfs:label` data.
- 533 of 676 available SPARQL endpoints contain at least 10 or more comment or label objects. Those containing less than 10 are excluded.
- 435 of 676 available SPARQL endpoints include more than 1.000 words, the rest of them are excluded.
- 77 of 676 available SPARQL endpoints return more than 10.000 comment objects. Therefore, sampled only the first 10.000 and ignored the rest.
- 21.553.998 words are extracted in total from these labels and comments.

The distribution of label and comment usage in the remaining 533 endpoints are depicted in Fig. 1. Almost half of the endpoints contain more than 8.192 labels and 15% of the endpoints contain more than 8.192 comments. It should also be noted that 25% of the endpoints contain between 500–1.000 labels and comments.

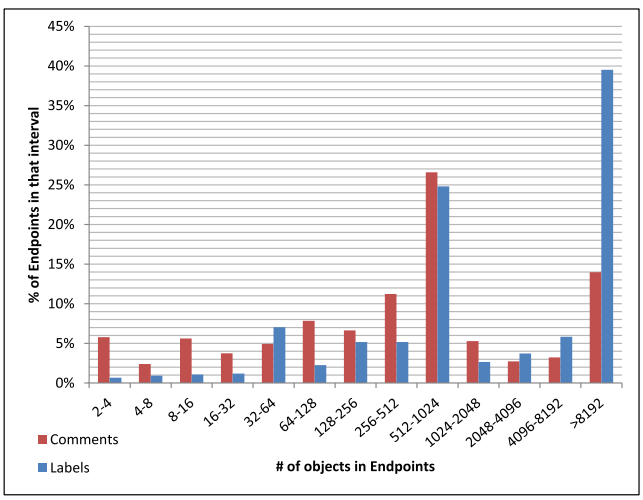

Fig. 1 Distribution of endpoints with the number of labels and comments that contain.

4.1 Topic and Tag Recommendation

The `Stfidf` and `Ctfidf` functions in Sect. 3.3 are applied on the data set with both hypernym and topic semantics term frequencies separately. Top scoring hypernym (h) and topics (t) both for `comment` (c) and `label` (l) words are sampled in Table 1. The sample data in this table contains the endpoints having more than 100K triples together with their top scoring words and their semantic terms (extracted by using WordNet). All words and terms (totally 858,815 records) are used in the classification tasks performed as explained in Sect. 4.2. The usage of these terms in the classification algorithms may increase the accuracy of the classification tasks, which are discussed further in Sect. 4.3. The complete list of these records is available for further research and analysis in the project repository[†]. Whereas this study focuses only on the classification of SPARQL endpoints by using topic and hypernyms, a topic term can also be used as a topic recommendation and a hypernym term can be used as a tag recommendation for the SPARQL endpoints. (e.g. well known endpoints *#20 biordf.net* has the c_t term "biochemistry", *#24 rdf.imim.es* has the c_t term "molecular biology" and the c_h term "drug")

4.2 Classification of Linked Data Sources

In order to measure the effect of the proposed tfidf scoring on classification tasks, Stfidf and Ctfidf scoring functions are applied on the document vectors (created from Linked Data sources as explained in Sect. 3.4). These data sources are then classified and tested by using 7 different classification techniques as listed in Sect. 3.4.

From Fig. 2 to Fig. 5, the maximum and average accuracy results of different classifiers and scoring methods are illustrated. In the graphs the following list of abbreviations are used in legends. For example, c_h_ctf stands for rdf:comments, WordNet hypernyms and Ctfidf scoring are being used.

- c: rdf:comment
- l: rdf:label
- h: WordNet Hypernym
- t: Wordnet Topic
- tf: tfidf score
- stf: Stfidf score
- ctf: Ctfidf score
- lvl: Wordnet Second Level terms

In Fig. 2, both Ctfidf and Stfidf scoring are applied before running different classifiers on `comments`. The accuracy results are calculated both for `Wordnet Hypernyms` and `Wordnet Topics`. According to this figure, the semantic scoring significantly increases the accuracy results compared to standard tfidf scoring. With the hypernym parameter and Stfidf scoring (c_h_stf), the accuracy increases up to 80% for the Naive Bayes classifier.

[†]github.com/semihyumusak/SparqlEndpointClassification

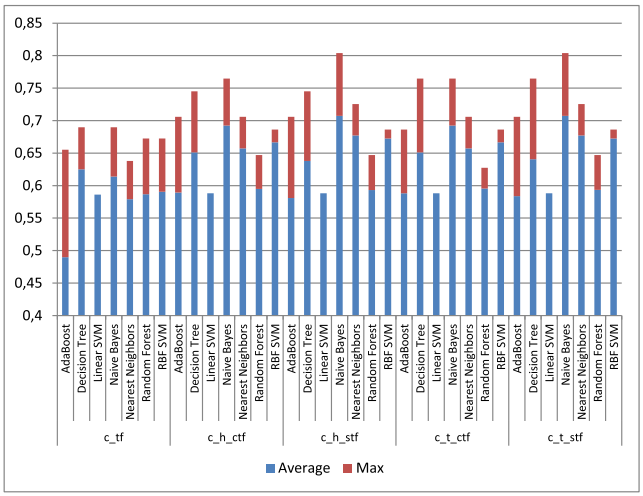

Fig. 2 Accuracy of classification methods and scoring methods on comment semantics

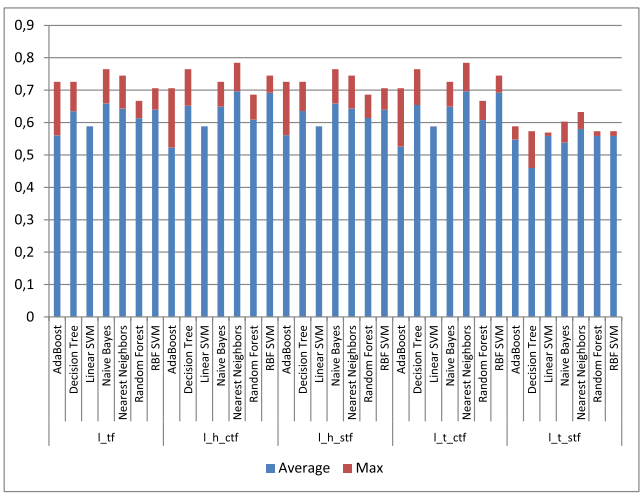

Fig. 3 Accuracy of classification methods and scoring methods on label semantics

In Fig. 3, both Ctfidf and Stfidf scoring are applied before running different classifiers on `labels`. The accuracy results are also calculated both for `Wordnet Hypernyms` and `Wordnet Topics`. According to this figure, whereas the semantic scoring results in 78% (l_h_ctf) accuracy for Nearest Neighbors classifier, tfidf scoring results in 73% accuracy for the same algorithm and 76% in Naive Bayes classifier.

In Fig. 4, both Ctfidf and Stfidf scoring are applied before running different classifiers on `comments`. The accuracy results are calculated both for `Wordnet Second Level Hypernyms` and `Wordnet Second Level Topics`. According to this figure, the semantic scoring results in a maximum accuracy of 80% (Naive Bayes c_h_stf_lvl), whereas tfidf score results in a maximum accuracy of 69% (Naive Bayes c_tf). The accuracy results of Fig. 2 and Fig. 4 are further statistically analyzed in Sect. 4.3.1 and Sect. 4.3.2.

Table 1 Top Stfidf scored terms extracted from the SPARQL Endpoints

		l_t		l_h		c_t		c_h	
#	Endpoint	word	term	word	term	word	term	word	term
1	digital-agenda-data.eu/sparql	reviews	accounting	reviews	accounting	quarter	professiona.	indicator	coloring ma.
2	lab.environment.data.gov.au/sparql	pilot	aircraft	horn	noisemaker	stations	navy	percentile	mark
3	newt.oerc.ox.ac.uk:8890/sparql	cultures	archeology	contrast	scope	record	photography	strains	nervousness
4	linked.opendata.cz/sparql	adenosine	biochemistry	bpi	density	net	field hockey	exchange	capture
5	glycoinfo.org/lodestar/sparql	rna	biochemistry	argument	computer ad.	charge	psychoanaly.	affinity	kinship
6	sparql.wikipathways.org	rna	biochemistry	hells	imaginary p.	set	psychology	set	abstraction
7	data.bbib.no/sparql	posting	bookkeeping	contraindic.	reason	synagogue	Judaism	synagogue	place of wo.
8	data.logainm.ie/sparql	posting	bookkeeping	constituent	syntagma	bishop	Roman Catho.	guardhouse	headquarters
9	location.testproject.eu/sparql	decision	boxing	decision	result	port	ship	exchange	capture
10	cpsv.testproject.eu/sparql	decision	boxing	decision	result	port	ship	exchange	capture
11	ieeevis.tw.rpi.edu/sparql	bidding	bridge	serf	thrall	sides	animal	citation	speech act
12	cr.eionet.europa.eu/sparql	renting	car	growing	production	filling	dentistry	disclaimer	repudiation
13	semantic.eea.europa.eu/sparql	renting	car	renting	transaction	tack	seafaring	pentecost	Jewish holy.
14	linkedstat.spaziodati.eu/sparql	book	card game	section	expanse	processor	photography	articles	determiner
15	lod.sztaki.hu/sparql	book	card game	email	electronic.	charge	tax	charge	liabilities
16	dbpedia-live.openlinksw.com/sparql	grace	Christian t.	philosopher	scholar	hero	Greek mytho.	dip	angle
17	live.dbpedia.org/sparql	grace	Christian t.	philosopher	scholar	hero	Greek mytho.	dip	angle
18	data.oceandrilling.org/sparql	grace	Christian t.	jacobs	patriarch	charge	tax	drill	training
19	sadiframework.org/registry/sparql	accession	civil law	citations	speech act	rna	biochemistry	insert	break
20	biordf.net/sparql	accession	civil law	citations	speech act	rna	biochemistry	insert	break
21	data.utpl.edu.ec/utpl/lod/sparql	ambrosia	classical m.	amphisbaena	mythical mo.	ambrosia	classical m.	inclination	angle
22	serendipity.utpl.edu.ec/lod/sparql	ambrosia	classical m.	amphisbaena	mythical mo.	ambrosia	classical m.	inclination	angle
23	data.cubiss.nl/sparql	alcides	classical m.	punt	kick	accounts	history	accounts	record
24	rdf.imim.es/sparql	article	contract	article	determiner	gene	molecular b.	medicine	drug
25	sparql.openmobilenetwork.org	article	contract	subjects	term	bengali	Hinduism	bengali	Asian
26	data.linkedu.eu/ocw/query	cmb	cosmology	terrorists	radical	margin	corporate f.	factorizati.	resolution
27	fantom5.nanopub.org/sparql	inversions	counterpoint	insert	break	rna	biochemistry	insert	break
28	rdf.neuinfo.org/sparql	inversions	counterpoint	insert	break	editing	literature	recombinati.	combining
29	wiktionary.dbpedia.org/sparql	drop	drug	check	chess move	station	navy	sentence	final judgm.
30	mlode.nlp2rdf.org/sparql	drop	drug	check	chess move	station	navy	sentence	final judgm.
31	wiktionary.dbpedia.org/sparql	drop	drug	check	chess move	station	navy	sentence	final judgm.
32	linked-statistics.gr/sparql	community	ecology	code	coding syst.	divisions	botany	citizenship	legal status
33	en.openei.org/sparql	utilities	economics	easements	prerogative	energy	physics	waste	deed
34	data.aalto.fi/sparql	consumption	economics	fatigue	duty assign.	disturbances	psychiatry	truss	bandage
35	aliada.scanbit.net:8890/sparql	use	economics	carriers	immune	judith	Apocrypha	manifestati.	protest
36	bis.270a.info/sparql	education	education	subjects	term	quarter	professiona.	clauses	grammatical.
37	fao.270a.info/sparql	education	education	subjects	term	quarters	professiona.	clauses	grammatical.
38	imf.270a.info/sparql	education	education	subjects	term	quarters	professiona.	clauses	grammatical.
39	sparql.asn.desire2learn.com:8890/sparql	education	education	correlation	parametric.	habits	religion	expectation	mean
40	data.webfoundation.org/sparql	education	education	computers	machine	relationship	anthropology	rectificati.	refining
41	bfs.270a.info/sparql	education	education	subjects	term	quarters	professiona.	clauses	grammatical.
42	stats.270a.info/sparql	repeaters	electrical.	parities	bit	xxx	genetics	xxx	sex chromos.
43	services.data.gov.uk/statistics/sparql	vicars	Episcopal C.	vicars	clergyman	councils	Christianity	councils	assembly
44	linkeddata.ge.imati.cnr.it:8890/sparql	cultivation	farming	accretion	increment	relationship	anthropology	curry	dish
45	epo.publicdata.eu/sparql	foils	fencing	vibrations	wave	work	physics	conviction	final judgm.
46	dbpedia.inria.fr/sparql	vampire	folklore	bengali	Asian	sabre	fencing	confession	penance
47	wordnet.okfn.gr:8890/sparql	americana	furniture	americana	artifact	record	photography	head	coil
48	wordnet.okfn.gr:8890/sparql	americana	furniture	americana	artifact	record	photography	head	coil
49	open-data.europa.eu/sparqlep	games	game	exchange	capture	telecommuni.	telecom	indicator	coloring ma.
50	leipzig-data.de:8890/sparql	games	game	passage	legislation	frau	German	frau	title of re.
51	linguistic.linkeddata.es/sparql	translation	genetics	translation	transformat.	costas	vertebrate	costas	bone
52	lod.nature.go.kr/sparql	characters	genetics	characters	attribute	phylum	biology	phylum	social group
53	data.linkedtv.eu:8890/sparql	herr	German	mensch	good person	exhibitions	art	plate	base
54	nl.dbpedia.org/sparql	athene	Greek mytho.	bolt	abandonment	moses	Old Testame.	libel	defamation
55	data.metamatter.nl/sparql	rabbi	Hebrew	showrooms	panopticon	church	church serv.	hackers	programmer
56	hanne.aksw.org:8892/sparql	rabbis	Hebrew	fighter	airplane	baldr	Norse mytho.	comet	extraterres.
57	ichoose.tw.rpi.edu/sparql	body	homo	column	file	charge	tax	charge	liabilities
58	data.clarosnet.org/sparql	mihrab	Islam	amphitheater	gallery	fathers	Christianity	fathers	theologian
59	data.allie.dbcls.jp/sparql	antigen	immunology	mp	lawman	processor	photography	processor	worker
60	zbw.eu/beta/sparql/stw/query	industries	industry	inflation	explosion	range	mathematics	connections	supplier
61	zbw.eu/beta/sparql/stw/query	industries	industry	inflation	explosion	range	mathematics	connections	supplier
62	services.data.gov.uk/education/sparql	menorah	Judaism	menorah	candelabrum	insert	film	insert	break
63	smartcity.linkeddata.es/sparql	temple	Judaism	mess	dining room	council	Christianity	pilot	aviator
64	environment.data.gov.uk/sparql/bwq/	shore	lake	birling	twirl	station	navy	colonies	animal group
65	data.globalchange.gov/sparql	hybrid	Latin	desktop	screen	literature	literature	manifestati.	protest
66	cs.dbpedia.org/sparql	tristan	legend	relativity	scientific.	hymen	Greek mytho.	offside	mistake
67	dati.san.beniculturali.it/sparql	justice	legislation	curia	administrat.	processor	photography	processor	worker
68	db.lodc.jp/sparql	circulation	library sci.	immunity	condition	temples	Judaism	ceramics	instrumenta.
69	data.linkedu.eu/kis/query	literature	literature	quartile	mark	completion	American fo.	quartile	mark
70	lod.bco-dmo.org/sparql	transmitter	microorgani.	squid	seafood	crown	dentistry	parameter	computer ad.
71	sparql.hegroup.org/sparql	nodule	mineralogy	tonicity	tension	rna	biochemistry	spasms	constriction
72	linkeddata.uriburner.com/sparql	smash	motor vehic.	disclaimer	repudiation	television	television	routers	device
73	eatld.et.tu-dresden.de/sparql	connections	narcotic	connections	supplier	accounts	history	devices	emblem
74	pubmed.bio2rdf.org/sparql	neurology	neurology	neurology	medical spe.	processor	photography	processor	worker
75	proxy.uriburner.com/sparql	ninja	Nipponese	raises	gamble	television	television	routers	device
76	uriburner.com/sparql	ninja	Nipponese	raises	gamble	television	television	routers	device
77	wit.istc.cnr.it:8894/sparql	joseph	Old Testame.	piste	ski run	brother	religion	advocate	lawyer
78	healthdata.tw.rpi.edu/sparql	lot	Old Testame.	indicators	coloring ma.	channels	river	column	file
79	wit.istc.cnr.it:8894/sparql	joseph	Old Testame.	piste	ski run	brother	religion	advocate	lawyer
80	data.ox.ac.uk/sparql	mover	order	appointments	disposal	literature	literature	fullerenes	carbon
81	es-la.dbpedia.org/sparql	aves	ornithology	phylum	social group	ishmael	Old Testame.	umma	community
82	bioportal.bio2rdf.org/sparql	fenestra	otology	apophysis	outgrowth	cytochrome	biochemistry	epitopes	situation
83	ruian.linked.opendata.cz/sparql	processor	photography	processor	hardware	processor	photography	processor	worker
84	kaiko.getalp.org/sparql	processor	photography	processor	hardware	processor	photography	processor	worker
85	crashmap.okfn.gr:8890/sparql	processor	photography	driver	utility pro.	processor	photography	processor	worker
86	virtuoso.gbpn.org/sparql	energy	physics	energy	physical ph.	energy	physics	energy	physical ph.
87	linkedspl.bio2rdf.org/sparql	ringers	quoits	balm	remedy	insert	film	insert	break
88	linkeddata.finki.ukim.mk/sparql	ringer	quoits	infusion	instillation	diana	Roman mytho.	infusion	instillation
89	matvocab.org/sparql	tracer	radiology	accelerator	activator	pitch	ship	reinforceme.	stimulation
90	eu.dbpedia.org/sparql	mars	Roman mytho.	camp	military qu.	amazon	Greek mytho.	frau	title of re.
91	linked-data.org/sparql	optative	Sanskrit	clauses	grammatical.	tao	Taoism	abaya	robe
92	opendata-bundestag.de/sparql	optative	Sanskrit	clauses	grammatical.	tao	Taoism	abaya	robe
93	semanticlab.jrc.ec.europa.eu:4433/sparql	veda	Sanskrit	veda	sacred text	axiom	logic	ontology	arrangement
94	lodlaundromat.org/sparql	header	soccer	nodes	point	header	soccer	client	computer
95	sparql.backend.lodlaundromat.org	header	soccer	nodes	point	header	soccer	client	computer
96	dati.camera.it/sparql	dona	Spanish	mafia	organized c.	account	history	ai	agency
97	data.sepa.org.uk	don	Spanish	bail	legal system	water	river	water	thing
98	spcdata.digitpa.gov.it:8899/sparql	don	Spanish	brig	penal insti.	pit	auto racing	tares	counterweig.
99	semantic.ckan.net/sparql	tag	tag	tag	touch	charge	tax	inflation	explosion
100	semantic.datahub.io/sparql	tag	tag	tag	touch	charge	tax	inflation	explosion
101	waes.servusnet.com/sparql	tag	tag	tag	touch	charge	tax	charge	liabilities
102	it.dbpedia.org/sparql	television	television	inclination	angle	anas	antiquity	salute	greeting
103	pt.dbpedia.org/sparql	television	television	fortes	volume	tv	television	tv	receiving s.
104	data.bnf.fr/sparql	nibelungen	Teuton	combats	battle	images	psychology	images	appearance
105	portal.chemicalsemantics.com/cs/sparql	h	thermodynam.	bond	recognizance	charges	tax	nucleus	midpoint
106	internal.opendata.cz:8890/sparql	zombie	voodoo	blitzkrieg	attack	net	field hockey	exchange	capture
107	id.dbpedia.org/sparql	obi	West Indies	jati	caste	jati	Hinduism	guru	religious l.
108	lod.gesis.org/thesoz/sparql	mensch	Yiddish	mensch	good person	tag	tag	tag	touch

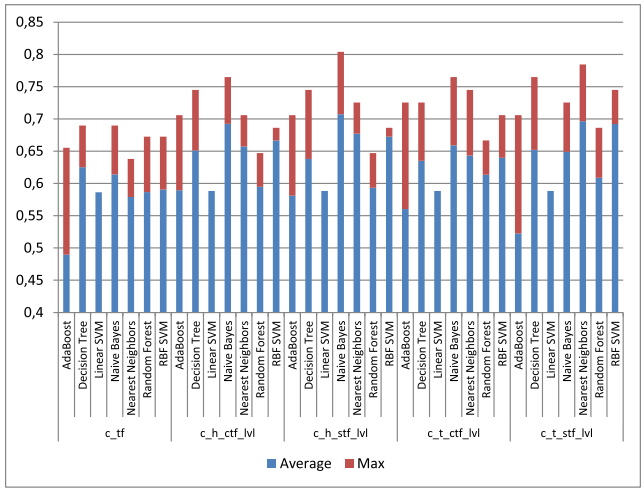

Fig. 4 Accuracy of classification methods and scoring methods on comments' second level semantics

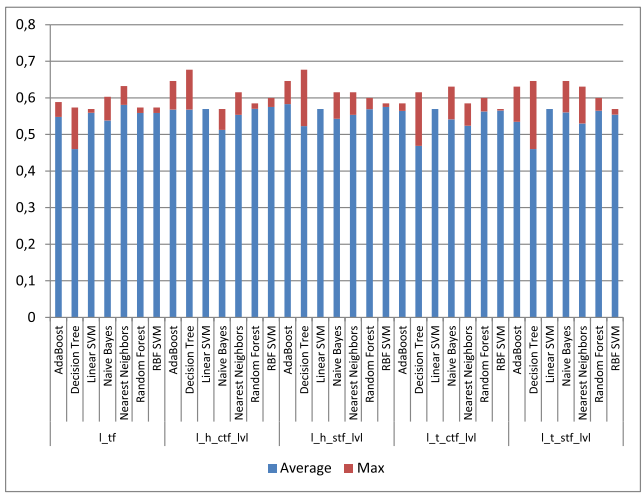

Fig. 5 Accuracy of classification methods and scoring methods on labels' second level semantics

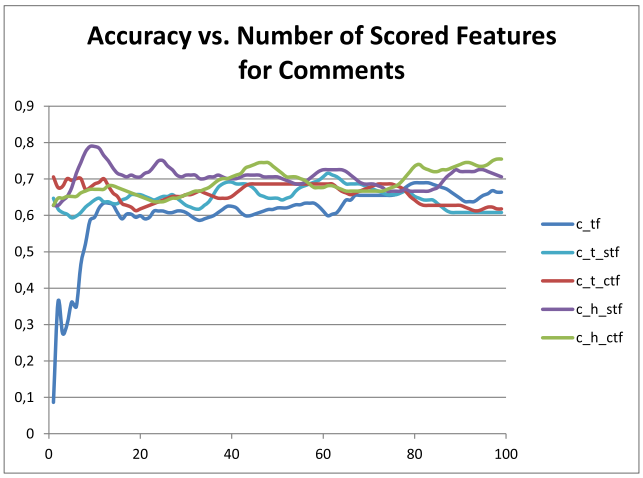

Fig. 6 Accuracy of Naive Bayes Classifier vs. number of comment semantics included based on the tf score (higher to lower score)

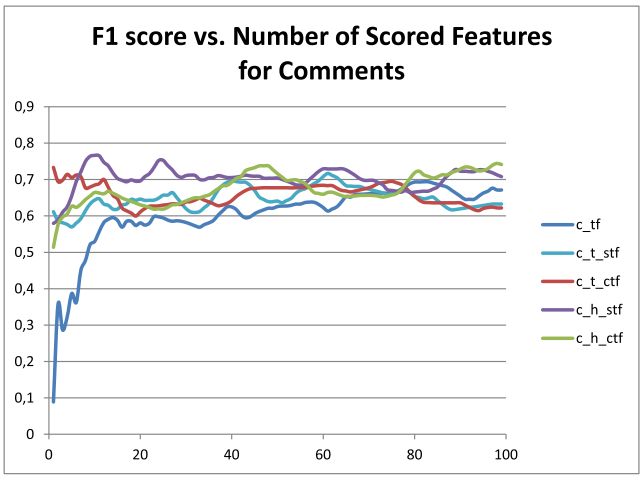

Fig. 7 F1 Score of Naive Bayes Classifier vs. number of comment semantics included based on the tf score (higher to lower score)

In Fig. 5, both Ctfidf and Stfidf scoring are applied before running different classifiers on `labels`. The accuracy results are calculated both for `Wordnet Second Level Hypernyms` and `Wordnet Second Level Topics`. According to this figure, the accuracy increases to 68% for semantic scoring (Decision Tree l_h_stf_lvl and l_h_ctf_lvl), whereas tfidf score results in a maximum score of 63% (Nearest Neigbors l_tf). The accuracy results of Fig. 3 and Fig. 5 are further statistically analyzed in Sect. 4.3.1 and Sect. 4.3.2.

As explained above, Figs. 2 to 5 summarize the effect of different inputs and scoring methods using different classifiers. Based on the results, Naive Bayes classifier performs better than the other classifiers in most of the cases. To examine the effect of the number of features scored by Stfidf and Ctfidf methods, a detailed illustration of the accuracy and F1 scores are drawn for the Naive Bayes classifier. These graphs are created by using an incremental feature selection method, which is scored by Stfidf and Ctfidf scores. In each of these experiments, the classification results are calculated for the top 100 features.

According to Figs. 6 and 7, standard tfidf scoring on comments results in lower accuracy and F1 scores, whereas hypernym and topic enhanced semantic scoring results in higher scores. On the other hand, the classification results by using labels as features do not show significant changes between scoring methods. The textual content under label properties are usually shorter than comment properties. Due to the lack of descriptive longer sentences, labels are not seen as a good feature for the classification task. However, it should be noted that, there is a difference in the topic related Stfidf (l_t_stf) classification results between 20–40 top features in Figs. 8 and 9.

4.3 Statistical Analysis of the Results

In this section, the prediction accuracy scores are analyzed

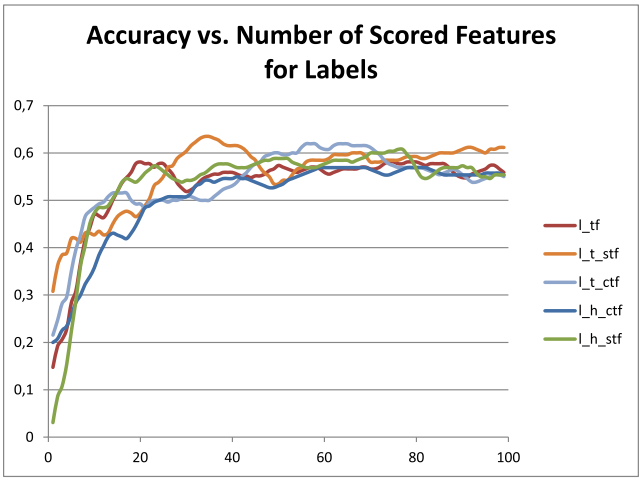

Fig. 8 Accuracy of Naive Bayes Classifier vs. number of label semantics included based on the tf score (higher to lower score)

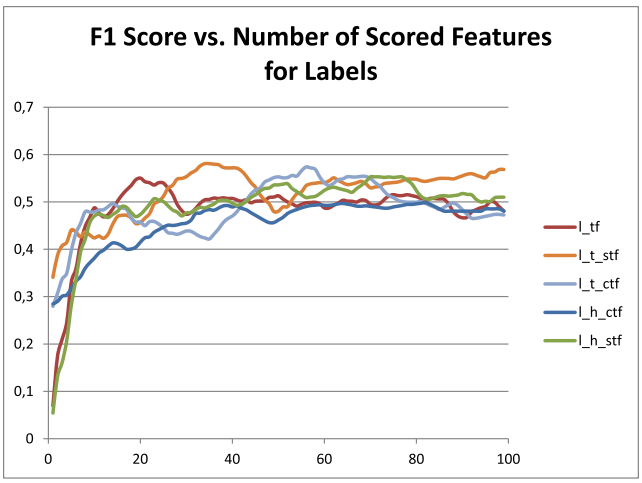

Fig. 9 F1 score of Naive Bayes Classifier vs. number of label semantics included based on the tf score (higher to lower score)

to identify the statistical significance of the accuracy difference between scoring methods. In order to perform a significant difference analysis, Kruskal-Wallis H test [14] and Mann-Whitney U test [18] are applied on the prediction accuracy scores. The Kruskal-Wallis H test is explained as "a rank-based nonparametric test that can be used to determine if there are statistically significant differences between two or more groups of an independent variable. It extends the Mann-Whitney U test when there are more than two groups." [1]. The use of the Mann-Whitney U test is defined as to "compare differences between two independent groups" [2]. These tests are applied first on Average accuracy scores in Sect. 4.3.1, then on Maximum accuracy scores in Sect. 4.3.2.

4.3.1 Analysis on Average Prediction Accuracy Scores

In order to perform a significant difference analysis, Kruskal-Wallis H test is applied on the average accuracy scores. Then, Mann-Whitney U test is applied to identify the source of the difference, which makes possible the binary analysis between each method. In Table 2, based on the mean ranks, Ctfidf accuracy results are higher than Stfidf and tfidf accuracy scores. According to the Kruskal-Wallis H test result, there is a significant difference between different scoring methods ($p < 0,05$). In order to identify the source of the significant difference by comparing binary groups, Mann-Whitney U test is applied and the results are tabulated as below.

Table 2 Results of the Kruskal-Wallis H test on the mean accuracy values in terms of different scoring methods (tfidf, Stfidf, Ctfidf)

Method	N	Mean Rank	df	chi-square	p
tfidf	56	70,64285714	2	6,853863393	*0,032
Stfidf	56	90,58928571			
Ctfidf	56	92,26785714			
Total	168				

* Statistical significance at $p < 0,05$

Table 3 Binary comparisons of the mean accuracy values between the scoring methods by using the Mann-Whitney U test

Methods	N	Mean Rank	Sum of Ranks	U	p
tfidf	56	50,50	2828,00	1232	*0,05
Stfidf	56	62,50	3500,00		
tfidf	56	48,64	2724,00	1128	*0,01
Ctfidf	56	64,36	3604,00		
Stfidf	56	56,59	3169,00	1563	0,98
Ctfidf	56	56,41	3159,00		

* Statistical significance at $p < 0,05$

Table 4 Binary comparisons of the average accuracy values between first level and second level semantics by using the Mann-Whitney U test

Semantics	N	Mean Rank	Sum of Ranks	U	p
1st Level	56	56,93	3188,00	1544	0,89
2nd Level	56	56,07	3140,00		

* Statistical significance at $p < 0,05$

In Table 3, binary comparison results are listed. According to this table, Stfidf scoring results is significantly higher than tfidf scoring results ($U = 1232$; $p = 0,05$; $p < 0,05$). Between Ctfidf and tfidf scoring methods, there is a significant difference in favor of the Ctfidf score ($U = 1128$; $p = 0,01$; $p < 0,05$). Nevertheless, there is no significant difference between Stfidf and Ctfidf scores ($U = 1563$; $p = 0,98$; $p > 0,05$).

In order to analyze the statistical significance of the average accuracy scores between first level semantics and second level semantics, Mann-Whitney U test is additionally applied on the accuracy results by considering the level as the independent variable. The results are listed in Table 4.

According to the results listed in Table 4, there is no significant average accuracy difference between the first level semantics and second level semantics ($U = 1544$; $p = 0,89$; $p > 0,05$).

4.3.2 Analysis on Maximum Prediction Accuracy Scores

Similar to the previous section, Kruskal-Wallis H test is also

Table 5 Results of the Kruskal-Wallis H test on the max. accuracy values in terms of different scoring methods (tfidf, Stfidf, Ctfidf)

Method	N	Mean Rank	df	chi-square	p
tfidf	56	70,5	2	7,885	*0,019
Stfidf	56	95,86607143			
Ctfidf	56	87,13392857			
Total	168				

* Statistical significance at $p < 0,05$

Table 6 Binary comparisons of the max. accuracy values between the scoring methods by using the Mann-Whitney U test

Methods	N	Mean Rank	Sum of Ranks	U	p
tfidf	56	48,1428571	2696	1100	*0,006
Stfidf	56	64,8571429	3632		
tfidf	56	50,8571429	2848	1252	0,065
Ctfidf	56	62,1428571	3480		
Stfidf	56	59,5089286	3332,5	1399	0.326
Ctfidf	56	53,4910714	2995,5		

* Statistical significance at $p < 0,05$

Table 7 Binary comparisons of the max. accuracy values between first level and second level semantics by using the Mann-Whitney U test

Level	N	Mean Rank	Sum of Ranks	U	p
1st Level	56	56,57	3168,00	1564	0,98
2nd Level	56	56,43	3160,00		

* Statistical significance at $p < 0,05$

applied on the maximum accuracy scores. Mann-Whitney U test is also applied to identify the source of the difference. In Table 5, based on the mean ranks, Stfidf accuracy results are higher than Ctfidf and tfidf scores respectively. According to the Kruskal-Wallis test, which is applied to understand whether there is a significant difference between the groups, there is a significant difference between different scoring methods. In order to identify the source of the significant difference by comparing binary groups, Mann-Whitney U test is applied and the results are tabulated as below.

In Table 6, Stfidf scoring results is significantly higher than tfidf scoring results ($U = 1100$; $p = 0,01$; $p < 0,05$). Between Ctfidf and tfidf scoring methods, there is no significant difference ($U = 1252$; $p = 0,065$; $p > 0,05$). Similarly, there is no significant difference between Stfidf and Ctfidf scores ($U = 1399$; $p = 0,326$; $p > 0,05$).

In order to analyze the statistical significance of the average accuracy scores according to first level semantics and second level semantics, Mann-Whitney U test is additionally applied on the accuracy results by considering the level as the independent variable. The results are listed in Table 7.

According to the results listed in Table 7, there is no significant average accuracy difference between the first level semantics and second level semantics ($U = 1564$; $p = 0,98$; $p > 0,05$).

5. Conclusion

In this study, linked data sources are assumed to be single documents, which include many sentences to be experimented for a topical classification method. As linked data sources contain textual information as properties of their graph nodes, the textual properties (rdfs:comment and rdfs:label) in those sources was used to create an explanatory text document. These documents were used as the training and test sets of the classification algorithms. While analyzing those extracted documents, we have used WordNet to semantically enhance the feature vectors. As it is explained in this paper, the topic and hypernymy related keywords may create significant differences in some conditions (discussed in Sect. 4.3) on the prediction accuracy scores of the classification algorithms when used together with a semantic scoring function proposed in this paper. As explained in Sect. 4.3, the difference between Ctfidf and Stfidf scoring methods are not significant; however, both of these methods performs significantly better than standard tfidf scoring method. On one hand, semantic scoring provides an ordered list of topics and hypernym terms, which can be used as a recommender system for topic and tag recommendations. On the other hand, these scores can be used to improve a classification algorithm, which can be used to classify a newly discovered linked data source. The proposed scoring methodology is developed and shared as a public repository. All source code developed during this data collection, curation, and classification process can be accessed through the classification repository[†].

Acknowledgements

This research is supported by The Scientific and Technological research council of Turkey with grant number 1059B141500052 (Ref. No: B.14.2. TBT.0.06.01-21514107-020-155998).

References

[1] Kruskal-Wallis one-way analysis of variance, https://en.wikipedia.org/wiki/Kruskal-Wallis_one-way_analysis_of_variance
[2] Mann-Whitney U test, https://en.wikipedia.org/wiki/Mann-Whitney_U_test
[3] WordNet, https://en.wikipedia.org/wiki/WordNet
[4] M. Acosta, M.-E. Vidal, T. Lampo, J. Castillo, and E. Ruckhaus, ANAPSID: An adaptive query processing engine for SPARQL endpoints, In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol.7031, LNCS, pp.18–34, 2011.
[5] S. Auer, J. Demter, M. Martin, and J. Lehmann, "LODStats An Extensible Framework for High performance Dataset Analytics," In: Knowledge Engineering and Knowledge Management, vol.7603, pp.353–362, Springer Berlin Heidelberg, 2012.
[6] C.P. B, C. Binnig, E. Jim, W. May, D. Ritze, M.G. Skjæveland, A. Solimando, and E. Kharlamov, Detecting Similar Linked Datasets Using Topic Modelling, vol.9088, 2015, http://link.springer.com/10.1007/978-3-319-18818-8
[7] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," International Journal on Semantic Web and Information Systems, vol.5, no.3, pp.1–22, 2009.
[8] C. Buil-Aranda, A. Hogan, J. Umbrich, and P.Y. Vandenbussche, Sparql Endpoint Status, http://sparqles.okfn.org/

[†]github.com/semihyumusak/SparqlEndpointClassification

[9] R. Cyganiak and A. Jentzsch, The Linking Open Data cloud diagram, 2014, http://lod-cloud.net/
[10] A. Ferrara, D. Informatica, L. Genta, and S. Montanelli, "Linked Data Classification: a Feature-based Approach," Proceedings of the Joint EDBT/ICDT 2013 Workshops, pp.75–82, 2013, http://doi.acm.org.pitt.idm.oclc.org/10.1145/2457317.2457330
[11] O. Görlitz and S. Staab, SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions, COLD, 2011.
[12] O. Hartig, "SQUIN: a traversal based query execution system for the web of linked data," Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp.1081–1084, 2013, http://doi.acm.org/10.1145/2463676.2465231
[13] A. Jain and D. Zongker, Feature Selection: Evaluation, Application, and Small Sample Performance, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.19, no.2, pp.153–158, 1997, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=574797
[14] W.H. Kruskal and W.A. Wallis, "Use of ranks in one-criterion variance analysis," Journal of the American statistical Association, vol.47, no.260, pp.583–621, 1952.
[15] S. Lalithsena, P. Hitzler, A. Sheth, and P. Jain, Automatic Domain Identification for Linked Open Data, 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), pp.205–212, 2013, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6690016
[16] Y. Li, S.M. Chung, and J.D. Holt, "Text document clustering based on frequent word meaning sequences," Data & Knowledge Engineering, vol.64, no.1, pp.381–404, 2008.
[17] H. Liu and R. Setiono, "Incremental feature selection," Applied Intelligence, vol.9, no.3, pp.217–230, 1998.
[18] H.B. Mann and D.R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," Ann. Math. Statist., vol.18, no.1, pp.50–60, 03 1947, http://dx.doi.org/10.1214/aoms/1177730491
[19] R. Meusel and V. Sarca, "Towards automatic topical classification of LOD datasets," Linked Data on the Web at WWW Conference, 2015.
[20] G.A. Miller, "WordNet: a lexical database for English," Communications of the ACM, vol.38, no.11, pp.39–41, 1995.
[21] B. Mutlu, M. Mutlu, K. Oztoprak, and E. Dogdu, "Identifying trolls and determining terror awareness level in social networks using a scalable framework," In: 2016 IEEE International Conference on Big Data (Big Data), pp.1792–1798, Dec. 2016.
[22] K. Oztoprak, "Profiling subscribers according to their internet usage characteristics and behaviors," Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015, pp.1492–1499, 2015.
[23] K. Oztoprak, "Subscriber profiling for connection service providers by considering individuals and different timeframes," IEICE Transactions on Communications, vol.E99-B, no.6, pp.1353–1361, 2016, https://www.jstage.jst.go.jp/article/transcom/E99.B/6/E99.B_2015EBP3467/_article
[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol.12, pp.2825–2830, Oct. 2011.
[25] M. Saleem and A.-C.N. Ngomo, "HiBISCuS: Hypergraph-based source selection for SPARQL endpoint federation," The Semantic Web: Trends and Challenges, vol.8465, pp.176–191, Springer International Publishing, 2014.
[26] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Commun. ACM, vol.18, no.11, pp.613–620, 1975, http://doi.acm.org/10.1145/361219.361220
[27] U. Scaiella, D. Informatica, P. Ferragina, A. Marino, and M. Ciaramita, Topical Clustering of Search Results, Proceedings of the fifth ACM international conference on Web search and data mining (May), pp.223–232, 2012, http://dl.acm.org/citation.cfm?id=2124324
[28] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Comput. Surv., vol.34, no.1, pp.1–47, 2002, http://doi.acm.org/10.1145/505282.505283
[29] K.S. Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of documentation, vol.28, no.1, pp.11–21, 1972.
[30] S. Tuarob, L.C. Pouchard, P. Mitra, and C.L. Giles, "A generalized topic modeling approach for automatic document annotation," International Journal on Digital Libraries, http://link.springer.com/10.1007/s00799-015-0146-2, 2015.
[31] S. Yumusak, E. Dogdu, H. Kodaz, and A. Kamilaris, "SpEnD: Linked data SPARQL endpoints discovery using search engines," IEICE Trans. Inf. & Syst., vol.E100-D, no.4, pp.758–767, April 2017.

Semih Yumusak received the B.S. degree in Computer Engineering from Koc University in 2005 and MBA degree from Istanbul Bilgi University in 2008. He worked as a researcher in The Insight Centre for Data Analytics, Galway, Ireland in 2016. He is currently a PhD student in Computer Engineering at Selcuk University. His research interests include Semantic Web, Linked Data, Web Mining.

Erdogan Dogdu is a professor in the Computer Engineering Department at Cankaya University, Turkey. He received his BS degree in Computer Engineering and Science from Hacettepe University in 1987, MS and PhD degrees in Computer Science from Case Western Reserve University in 1992 and 1998 respectively. His recent research interests are in semantic web, web computing, big data analysis, and IoT.

Halife Kodaz graduated from Computer Engineering Department of Selcuk University with B.S. degree and M.S. degrees in 1999 and 2002, respectively. He received the Ph.D. degree in Electrical and Electronics Department from Selcuk University in 2008. He is an Associate Professor at the Computer Engineering Department at Selcuk University. His research interests are artificial intelligence and machine learning.