Information Sciences 426 (2018) 160–170

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Big Data and forensics: An innovative approach for a predictable jurisprudence

Massimiliano Giacalone[a,*], Carlo Cusatelli[b], Angelo Romano[c], Antonio Buondonno[c], Vito Santarcangelo[d]

[a] Department of Economics and Statistics, University of Naples 'Federico II', Italy
[b] Ionian Department, University of Bari "Aldo Moro", Italy
[c] iInformatica S.r.l.s., Corso Italia 77, Trapani, Italy
[d] Department of Mathematics and Computer Science, University of Catania, Italy

ARTICLE INFO

*Article history:*
Received 23 May 2017
Revised 13 October 2017
Accepted 15 October 2017
Available online 17 October 2017

*Keywords:*
Quality in law
Efficiency in law
Big data
Literal text similarity
Semantic text similarity

ABSTRACT

Nowadays, it is easy to trace a large amount of information on the web, to access documents and produce a digital storage.

The current work is submitted as an introduction to an innovative system for the investigation about notoriety of web data which is based on the evaluation of judicial sentences and it is implemented to reduce the duration of all processes.

This research also aims to open some new conjoint debates about the study and application of statistical and computational methods to web data on new forensics topics: text mining techniques enable us to obtain information which may be helpful to establish a statistical index in order to describe the quality and the efficiency in terms of law. It is also possible to develop an intelligent system about facts and judgments.



## 1. Introduction: The legal reason of our system

The statistical surveys highlight different purposes, and their fields of application are equally extensive. The legal one is particularly important for our work: applying the statistics to the final product of a trial – which is the sentence – we are given the chance to extract information about two essential aspects: the duration of a trial and the uniformity of judges' decision; these two elements are necessary in order to understand how efficient our judicial system can be and what the solution of a certain case would be [30].

Our main purpose is to create a computer system that – through a statistic analysis based on the legal databases stored on the web – examines the different decisions of the judges in such a way that we could obtain the following important data: the duration of a single trial, the solution adopted by a judge in a specific case and its correspondence with the other decisions of the judges who discussed the same case.

Before moving forward with the current analysis, it is important to under-line that an ordinary citizen, currently willing to know the Italian law, can use different legal databases [17].

The official decrees issued by the Constitutional Court are available for any consultation on the free websites http://www.cortecostituzionale.it and http://www.giurcost.it, while those issued by the Supreme Court can be found on http://www.iusexplorer.it.

* Corresponding author.
*E-mail address:* massimiliano.giacalone@unina.it (M. Giacalone).

https://doi.org/10.1016/j.ins.2017.10.036
0020-0255/

Among the benefits that we all could get from this new approach, there is the possibility to access the Italian law in a better and deeper way; moreover, our system, would allow the citizens to make a prognosis of their question before starting a trial simply by calculating the most shared opinion of the judges about a certain case. In case our system gives them a positive answer about the orientation of the Court on that question, the citizens will be encouraged to start the trial; while, in case the majority of the judges give a negative feedback on their request, they will be discouraged.

Therefore, the system also represents a way to reduce the burden of litigation and to fulfill, as a consequence, the important precept contained in the article 111 of the Italian Constitution where it is stated that every process must have a "reasonable length", and then, the European precept contained in the article 6 of the European Convention of the Human Rights which says that "everyone is entitled to a fair and public hearing within a reasonable time". The need for every citizen of a "predictable" jurisprudence has been established by both the first president of the Court of Cassation (in his annual relation about the administration of the justice) and by the authoritative doctrine that explicitly declared "the predictability can play even an economic role, since that if the decision is predictable, you can avoid going to the judge". The work seems to be coherent to the novelty recently introduced which gives free access to what has been stated by the Court of Cassation from 2009 onwards [18].

The paper is organized as follows: in Section 2 we show the Web data notoriety system state of the art and propose a new original weighted function to assign a notoriety score to the web sites. In Section 3 we evaluate the efficiency of the judicial systems looking at the data of the World Bank where the quality of the judicial processes index is emphasized. In paragraph 4 there are some important concepts on how to improve the notoriety system performance and, in the following section, an example to compare the different approaches is given. Afterwards, in Section 6, we describe the system proposed, the focus of the paper for the evaluation of the sentences and for the duration of the processes. Our idea has been tested using data provided by http://www.giustizia-amministrativa.it, because they are updated to 2017 and publicly available compared to the sites of civil justice and criminal justice in Italy. In sections 7 and 8, some conclusive notes and future developments are finally provided.

## 2. State of the art of web data notoriety system

The influence that the Web has on our society is constantly growing, then, it is not difficult to notice how it is becoming always more frequently the only way to reach a huge amount of data, to exchange opinions and increase our personal knowledge [8].

However, the data we are talking about do not always fulfill the unavoidable criteria of quality [38]. This problem can be analyzed through these two phenomena [5]: DISINFORMATION and MISINFORMATION. Disinformation meant as an intentional inaccurate or false information; misinformation with the meaning of an unintentional inaccurate information. Especially in these last years, a lot of disinformation tools were born and there is an interesting debate about these phenomena which starts from the reality that there is no tool or system able to protect the user by these two big issues and the common search engines return only the results of researches without measuring the quality of the data [4] Fig. 1.

The aim of a notoriety system, then, is to provide an automatic and intelligent score about the NOTORIETY of the data [24]; this can be imagined like an automatic controller that executes a normal web research in the following steps. The first three steps are the usual ones on which a common search engine is based: the users write what they are looking for and the engine provides the results through the crawler; these results are ranked by a text similarity logic and related to the popularity. The notoriety system provides a score $\phi$(xi) between 0 and 1 about text similarity thanks to a string comparison between the results and the input text [15].

This score shows an important feedback in order to understand how much 100 the results is near to the research. The second step consists in giving a notoriety score W(xi) to the source websites that reflects the reliability of the information

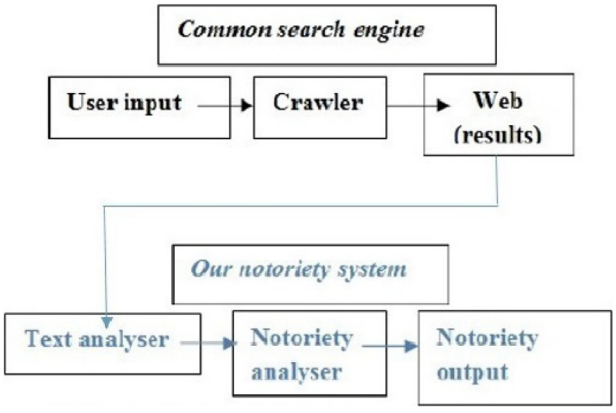



**Fig. 1.** First notoriety system.

| WEBSITE | NOTORIETY | APPLICATION FIELD |
|---|---|---|
| www.ansa.it | +3.0 | News |
| www.wikipedia.it | +2.0 | General |
| nonciclopedia.wikia.com | -3.0 | Funny |
| www.istruzione.it | +3.0 | Institutional |
| www.uniba.it | +3.0 | Institutional |
| www.sportnotizie24.it | +1.0 | SportNews |
| www.barzellette.net | -3.0 | Funny |
| www.bufale.net | -3.0 | Funny |
| www.altervista.org | -1.0 | Web Hosting |

**Fig. 2.** An extraction of notoriety knowledge-base.

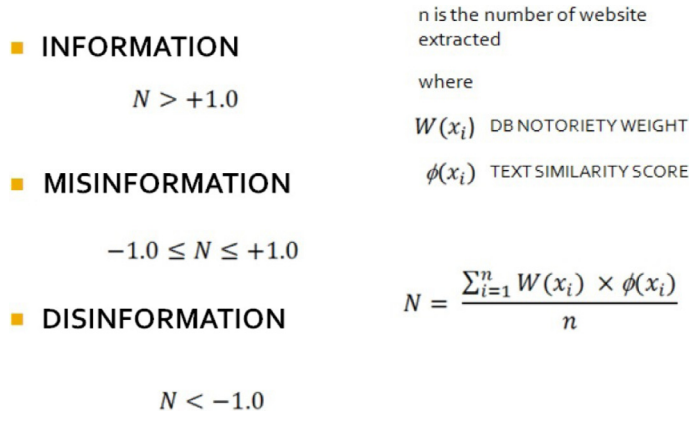



**Fig. 3.** Logic used for notoriety score.

source thanks to a notoriety database [20]. The previous image (Fig. 2) shows an extraction of notoriety database; considering that each notoriety and text similarity score can be defined N as total notoriety score for the input [7], the result can be read in this way: a score lower than −1 represents with high probability a disinformation, a result higher than +1 is supposed to be a reliable information, meanwhile a result between −1 and +1 probably shows an in-accurate information (misinformation) (Fig. 3).

This notoriety score of each website is between −3 (not reliable at all) and +3 (very reliable) and represents the notoriety of the information found; it is also related to the application field of the website if we consider a +3.0 score for institutional domain (e.g. .edu / .gov.it / .int / .museum). Then, the weight W(xi) gives a score on the notoriety of the website which is based on the logic that not all the sources have the same relevance. So we created a databases with almost 500 entries of famous website with a score from −3 to +3.

We assigned the point +3 to the government websites or Ansa or University, to all the websites that have a reliable and objective relevance. Instead, we gave +2 to official newspapers and +1 to all the websites of news or similar where the information could be influenced by a subjective opinion and so on; −1 for websites like Altervista, −2 to the websites of opinion (i.e. yahoo answer) and −3 to the tools that create explicit disinformation [40].

## 3. Quality of judicial process: A statistical analysis from OECD countries

The judiciary is an essential element in order to achieve a higher level of development: the satisfactory protection of rights generally encourages investment that earnings; meanwhile, an effective contract enforcement encourages business against opportunistic behavior and reduces transaction costs [37]. The duration of proceedings can be conceived as the interaction between requirement and offer of justice. Regarding the provide side, the typical factors are: quality and quantity related to financial and recruiting, to governance and organization of the courts, to the performance while using the assets (the latter concerns the degree of specialization, offices computerization and flows managing) [11]. The elements which influence the needs in terms of judicial system are: the expense of accessibility to its system, the spread of alternative

systems for dispute resolution, the particular legislation and the legal certainty [13]. Moreover, the lack of justice means economical costs, and the World Bank measures the phenomenon in its Doing Business report by means of the distance to frontier score (DTF) benchmarking economies with respect to regulatory best practice (showing the absolute distance to the best performance on each indicator) and the ease of doing business ranking (mutually comparing economies): those indicators are shown in Table 1 both from an overall and a more detailed viewpoint.

To calculate DTF score, each of the 36 component indicators $y_i$ (except for the total tax rate) is normalized to a common range $y_i^* = (worst_i - y_i) / (worst_i - frontier_i)$ where the frontier displays the best performance on the indicator across all the economies. Therefore, the overall distance to frontier score ODTF=$(\sum_{i=1}^{36} y_i^*/36) \times 100$ shows the distance, on a scale from 0 to 100, of each economy to the best performance observed on each of the indicators across all the economies. The ranking is determined by sorting DTF scores [14].

Furthermore, the World Bank measures time and cost for a commercial dispute through a first-instance court and the quality of judicial procedures index, evaluating whether a country has adopted procedures able to promote efficiency in the judicial system. The information comes from the analysis of their codes of civil procedure and other regulations as well as data from surveys completed by their lawyers and judges.

For instance, in Italy (Table 1) 1120 days are needed to solve a commercial dispute through a local first-instance court, more than doubling the average of high income OECD countries (538 days), placing Italy in the 111th place in the ranking of "enforcing contracts" (passed by only Slovenia and Greece), with a cost amounting to 23.1% of contract credit [32].

Finally, the quality of judicial processes index (QJPI), ranging from 0 to 18, measures whether each economy has been following a series of good practices in its judicial system in four areas: it's the sum of the scores on court structure and proceedings, case management, court automation (for example, in Italy this sub-indicator is equal to 3.0 out of 4.0 because the initial complaint can be electronically led through a particular platform within the competent court and it is possible to electronically realize a service of process for claims led before the competent court; moreover, court fees can be paid electronically within the competent court, but judgments rendered in commercial cases at all levels are not available for public use in publications of official gazettes, newspapers or on the internet (or court website). Alternative dispute resolution indexes place Italy at the beginning of the third quarter of the global distribution.

In order to stop this ineffectiveness, one of the key challenges is to use the web for the exchange of information between judicial operators: information and communication technology (ICT) is transforming the public sector, particularly as a part of the justice framework. The telematics civil procedure (TCP) is the set of regulations for the computerization of the civil trial: it particularly operates connecting lawyers and judicial workplaces [6].

The purpose of the TCP is to supplant the hard copy with the digital file through databases with subsequent time and costs savings, but it also aims to automate information and document flows allowing to formulate, sign and organize documents; to access to registers directly from desktop; to receive communications from the judicial staff; to make electronic payments [10]. In addition to this, it leads to the reduction of the vast amount of paper deposited in the courts. According to the latest data available at https://pst.giustizia.it/PST/resources/../PCT_Stato_arte_sintetico_31_05_2017.pdf, in Italy, the electronic filing of documents by lawyers and professionals between July 2016 and May 2017 reached an amount of 8,391,112 (12,39% more than previous year-period).

In the same period 19,054,106 communications and telematics notifications were delivered: the unit cost of a communication is estimated at around € 7, paying particular attention to the cost of the postal service and the average cost of the actions of the bailiff, and is applied as a precaution to 50% of telematics communication delivered, saving about 66 million euros.

Currently, there are over 6 million hits per day to the civil registers of all courts creating a very Big Data flow [36].

Nevertheless, statistics are not the only source to keep track of the judicial administration functioning: verdicts contain such an amount of information that, if properly extracted and organized, could provide analytical Big Data, which would be useful to validate the consistency of proofs or to infer investigative activities. This implicates semantic-based data extraction, development of technologies and implementation of case flow management" techniques ensuring work effectiveness: for example, the earlier examination and assignment of incoming processes to proceedings based to their features would identify problematic instances by reducing the length of all procedures [33]. ITC essentially facilitates knowledge and the exchange of data and information and is one of the main instruments of judicial innovation. The comparison between the approaches and experiences in various countries can facilitate the understanding of innovation processes and their peculiarities for the purpose of further quality improvements and cost to resolve debates in a more reasonable time.

## 4. Improvement of notoriety system performances

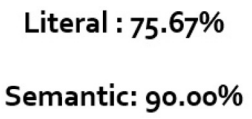
The approach presented in paragraph 2 represents the first example of notoriety system appeared in literature. Some of its limits are related to text similarity accuracy and problems about "notoriety website that writes about fake news" and weight accuracy [34].

### *4.1. Text similarity accuracy*

Text similarity is an important task that includes information retrieval, document clustering and, in our case, web scoring for understanding the quality of the information [19].

**Table 1**
"World Banks Doing Business 2016" rankings of high income OECD countries.

| | | | Enforcing contracts | | | | |
|---|---|---|---|---|---|---|---|
| Economy | Ease of doing business (overall rank) | Overall DTF | Rank | DTF | Time (Days) | Cost (% of claim) | Quality of judicial processes index (0–18) |
| Australia | 13 | 80.08 | 4 | 79.72 | 395 | 21.8 | 15.5 |
| Austria | 21 | 78.38 | 6 | 78.24 | 397 | 18.2 | 14.0 |
| Belgium | 43 | 72.50 | 53 | 64.25 | 505 | 18.0 | 8.0 |
| Canada | 14 | 80.07 | 49 | 65.49 | 570 | 22.3 | 10.5 |
| Chile | 48 | 71.49 | 56 | 62.81 | 480 | 28.6 | 9.0 |
| Czech Republic | 36 | 73.95 | 72 | 60.36 | 611 | 33.0 | 10.5 |
| Denmark | 3 | 84.40 | 37 | 68.56 | 410 | 23.3 | 10.0 |
| Estonia | 16 | 79.49 | 11 | 75.16 | 425 | 21.9 | 13.5 |
| Finland | 10 | 81.05 | 30 | 70.33 | 375 | 16.2 | 9.0 |
| France | 27 | 75.96 | 14 | 74.89 | 395 | 17.4 | 12.0 |
| Germany | 15 | 79.87 | 12 | 75.08 | 429 | 14.4 | 12.0 |
| Greece | 60 | 68.38 | 132 | 50.19 | 1580 | 14.4 | 12.0 |
| Hungary | 42 | 72.57 | 23 | 72.08 | 395 | 15.0 | 10.0 |
| Iceland | 19 | 78.93 | 35 | 69.10 | 417 | 9.0 | 7.5 |
| Ireland | 17 | 79.15 | 93 | 57.88 | 650 | 26.9 | 8.5 |
| Israel | 53 | 70.56 | 77 | 59.78 | 975 | 25.3 | 14.0 |
| *Italy* | *45* | *72.07* | *111* | *54.79* | *1120* | *23.1* | *13.0* |
| Japan | 34 | 74.72 | 51 | 65.26 | 360 | 23.4 | 7.5 |
| Korea, Rep. | 4 | 83.88 | 2 | 84.84 | 230 | 10.3 | 13.5 |
| Luxembourg | 61 | 68.31 | 17 | 73.32 | 321 | 9.7 | 8.5 |
| Netherlands | 28 | 75.94 | 91 | 58.09 | 514 | 23.9 | 6.0 |
| New Zealand | 2 | 86.79 | 15 | 74.25 | 216 | 27.2 | 11.0 |
| Norway | 9 | 81.61 | 8 | 77.14 | 280 | 9.9 | 10.0 |
| Poland | 25 | 76.45 | 55 | 63.44 | 685 | 19.4 | 10.5 |
| Portugal | 23 | 77.5 | 20 | 73.01 | 547 | 13.8 | 12.5 |
| Slovak Republic | 29 | 75.62 | 63 | 61.69 | 705 | 30.0 | 12.0 |
| Slovenia | 29 | 75.62 | 117 | 53.90 | 1160 | 12.7 | 11.0 |
| Spain | 33 | 74.86 | 39 | 67.63 | 510 | 18.5 | 10.0 |
| Sweden | 8 | 81.72 | 24 | 72.04 | 321 | 30.4 | 12.0 |
| Switzerland | 26 | 76.04 | 46 | 66.07 | 390 | 24.0 | 8.5 |
| United Kingdom | 6 | 82.46 | 33 | 69.36 | 437 | 43.9 | 15.0 |
| United States | 7 | 82.15 | 21 | 72.61 | 420 | 30.5 | 13.8 |




| Word1 | Word2 | Relation |
|---|---|---|
| Basil | Pesto | Correlation |
| Ocimum | Basil | Synonimous |

**Fig. 4.** Comparison of literal and semantic text similarity approach.

This weight is in a range from 0 (no similarity) to 1 (completely similarity). There are different approaches for making this comparing like literal approaches, semantic approaches or hybrid approaches [2].

The literal approach also called, string-based method, is the most simple method. It is based on a simple literary similarity between two words or in general two sentences or texts that are compared character by character. However, this is a very inaccurate method, as it doesn't get the concept hidden behind a sentence. Instead, semantic approach is the heart of the algorithms used in the modern search engine. Like suggest the name, this approach try to get the semantic relatedness between two concepts or words. It simulates the human way to understand if two terms have relatedness just analyzing the concept that they want to express [27].

These kind of methods are able to find out if two terms are synonymous and if two texts are talking about the same arguments but in different words [21]. In hybrid approach there is a combination of corpus and knowledge based and it is considered the best approach way to join good results in text similarity. In the last year a lot of efforts were made for improving the quality of semantic analysis and a lot of approach have been proposed [25].

The knowledge-based similarity is more efficient to give a reliable result, then, for this reason, in our work we have used a semantic thesaurus for a better TEXT SIMILARITY ANALYSIS. The previous image shows the differences between literal and semantic approach. (Fig. 4).

### 4.2. Parser fake control

The second limit is represented to the necessity of a controller of "notoriety website that writes about fake news". This limit has been solved by the introduction of a coefficient p, that is a "parser fake control". This coefficient can be 0 or 1 and it is used just for changing the sign of the result in the case of an high notoriety websites is talking about the information like a not true information in explicit way [35].

This is based on keywords that are able to recognize this possibility and change the result for getting a realistic result. Then, with this system it is possible to recognize with a realistic probability if an information on the web is reliable or not [12].

### 4.3. Weight accuracy

To improve the performance of the system has been modified the formula for the calculus of the notoriety score [31].

This representation considers the parser fake coefficient and the intensifier of high quality notoriety related to the high value of the two scores. This approach has been named "AVE" system and it is represented by the following images. (Figs. 5 and 6).

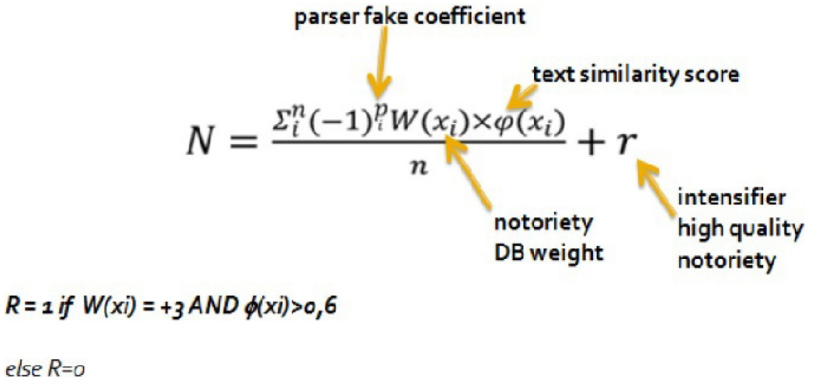



**Fig. 5.** "AVE" notoriety logic.

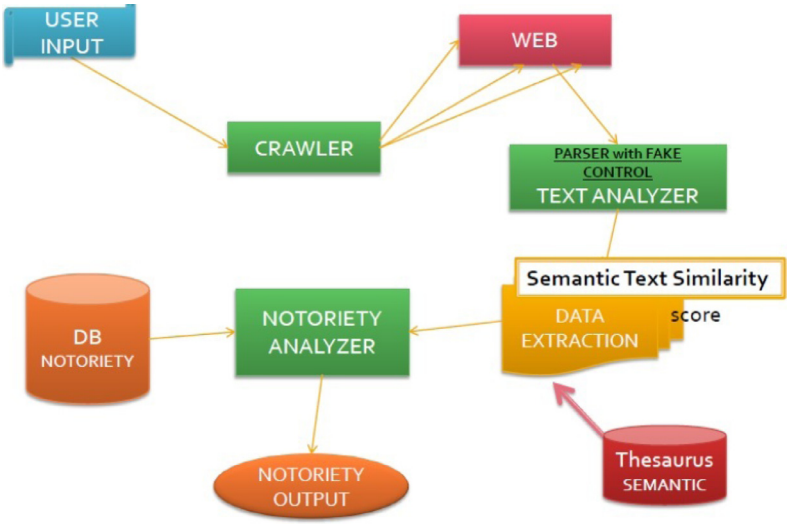



**Fig. 6.** "AVE" notoriety system.

## 5. Comparison between approaches

To better understand the logic of the "AVE" system, it is interesting to consider the analysis of the previous example ("snacks, ice cream and soft drinks toxic: the center of cancer asks maximum dissemination"), that we know is a false information. The following image (Fig. 7) show the comparison between the first notoriety system (literal), our "AVE" system in literal mode and our AVE system with semantic engine [26].

The "AVE" system (semantic) returns as result the value of $N = -1,66$ (Disinformation), the "AVE" literal returns $-1,58$ (Disinformation) and the first approach $-0,9$ (Misinformation). The accuracy of "AVE" system is better than the first approach.

The system proposed involves a Pre-processing module (Text Analyzer and Data Extraction Module) with the support of a Semantic Thesaurus and of an Analysis and Representation module (Notoriety Analyzer), supported, in turn, by the innovative "DB notoriety" as KB that follows an evolution of the approach proposed in [24]. It takes inspiration from the "High-level architecture for question answering" model which was highlighted in [23].

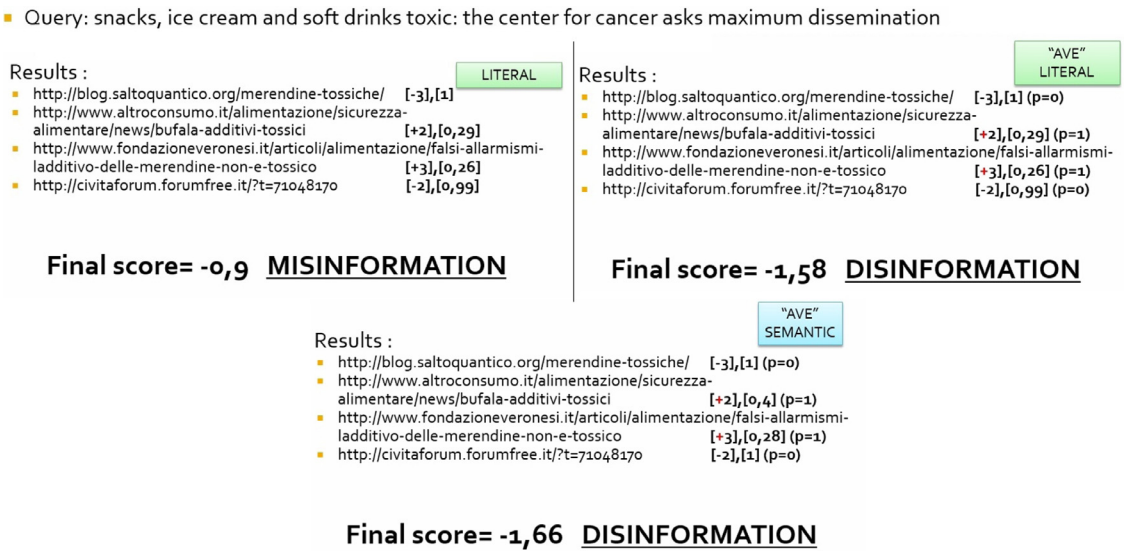



**Fig. 7.** Run of notoriety approaches.

## 6. System proposed for predictable jurisprudence

As state of the art, it is possible to find some studies about the automatic analysis of the sentences and it is a very important field of study in the statistical context [28].

In fact, a crawler, searching on the OSINT data, can extract from legal websites and examine -by means of a parser - the structure of the sentence. The sentence obtained can be segmented in two main parts: the "motivation" (that provides data and keywords of the process) and the "pqm" (that provides the sentence output). The keywords of the process are identified thanks to a "semantic knowledge base" [16] for a better text analysis. Thanks to the module of the "sentence segmentation", it is possible to extract the duration of the process considered, the keywords (input) of the contentious and the output decided by the judge [22].

The prototype system has been tested on our private repository of 100 sentences (called "SR - Sentences repository").

The module of sentence segmentation has performed correctly, producing a sentences thesaurus structured by duration, input (keywords) and sentence output.

Considering the size of sentences repository, the parameters of efficiency and quality in this moment cannot be representative. An important task to achieve is the extension of the repository.

To provide a better explanation of the logic of our system, we provide an example: the crawler retrieves sentences in the OSINT REPOSITORY (e.g. cortecostituzionale, giuricost, iusexplorer) and, from each sentence, composed by three parts (FACT, MOTIVATION and DEVICE or PQM), the crawler extracts dates, keywords (sentence input) and sentence output. The information retrieved is filtered by a semantic layer. The duration extracted and the information are recorded in the sentences thesaurus.

These information are, then, inserted in the sentences thesaurus" that is fundamental to the two indexes: efficiency (related to the duration of the process) and quality (related to the coherency of the sentences). Here we find two of our main aims: one for the user query (EVALUATION SYSTEM) and one for the automatic updating of the SENTENCES THESAURUS (as shown before) [39].

The first part is developed as an "user oriented" module by a parser/Text Analyzer (supported by a "semantic knowledge base"), and a core module (EVALUATION SYSTEM) related to the thesaurus (SENTENCE THESAURUS) that returns the sentence output. The user input can be a list of keywords or a text description as system query "eg. I killed the golden fish of my girlfriend". In the case of text description, the parser/text analyzer extracts the keywords of the user input, thanks to the joint analysis with the semantic knowledge base [29].

This KB is structured by words linked to each other through semantic relations. The key-words extracted become the input of the evaluation system that provides the query to the sentences thesaurus returning a sentence output with the relative score (eg. Absolution | Score: 98%) and citing the sentences related to that query. The user module aims also to achieve the target of sharing of information and awareness about Italian jurisprudence and the reduction of relative costs. The two modules presented can also be integrated in a single scheme that shows the logic flows of the automatic (on the right) and manual (on the left) parts (Fig. 8). We have implemented our system on the public dataset "http://www.giustizia-amministrativa.it".

This dataset is characterized by 2,208,139 sentences from 1908 to 2017.

A lot of sentences are not correctly available as information on the website or are not provided as html following our requirements, then, our parser has extracted a subgroup (Fig. 9). Considering the last 7 years (from 2010 to 2016), Table 2, the database repository is characterized by 1,225,754 sentences, that can be grouped for each year: We have produced a dataset called "MVA Sentence Dataset" of 1000 records, as extraction of sentences from 2010 to 2016 (the 0,08%), considering the needed requirements by our prototypal system (motivation and pqm) thanks from a parser analysis directly to the website. In fact, as the figure underlines, the parser has extracted the sentences from the repository and the sentence segmentation module extracts the duration of the process (from dates), the keywords (input) and the sentence output (p.q.m.).

Our system was developed with a new innovative approach after taking inspiration from the "High-level architecture for question answering" model which was highlighted in the paper entitled "A Survey on Question Answering Technology from an Information Retrieval Perspective" [23].

**Table 2**
http://www.giustizia-amministrativa.it, Number of sentences per year.

| Year | Number of sentences |
|---|---|
| 2010 | 238,759 |
| 2011 | 178,174 |
| 2012 | 210,840 |
| 2013 | 156,645 |
| 2014 | 156,286 |
| 2015 | 144,098 |
| 2016 | 140,951 |

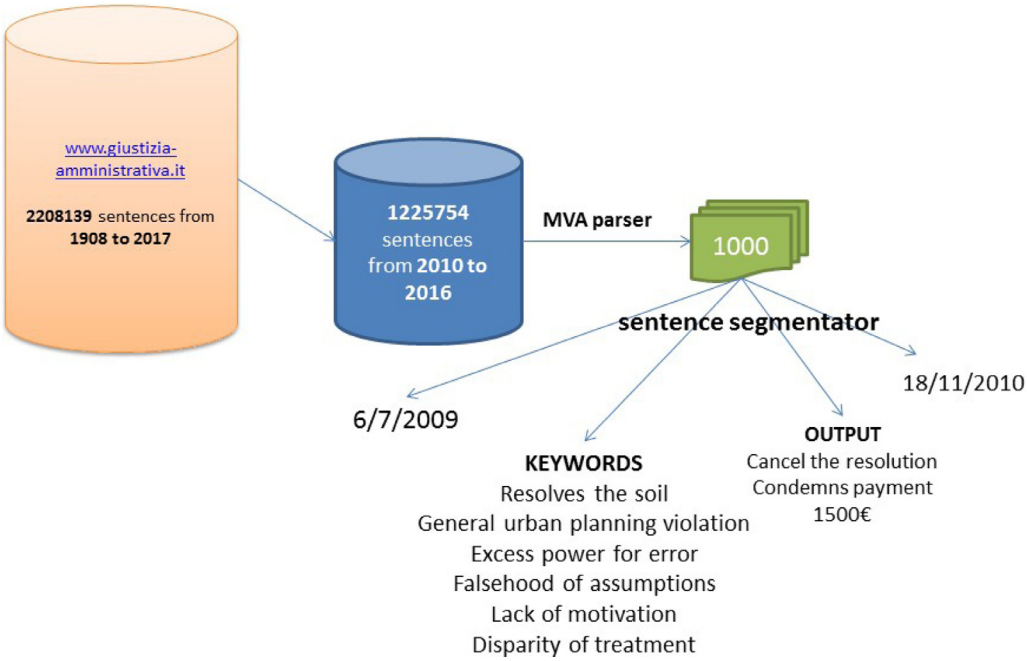



**Fig. 8.** Sentence system.

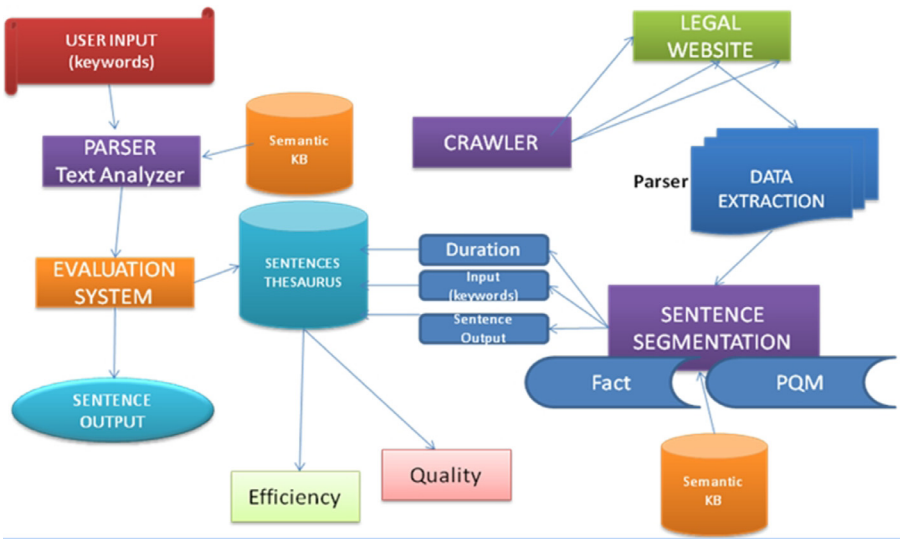



**Fig. 9.** Structure of our system.

Indeed, it uses a “Pre-processing module” (in our case, the Parser) and a module of “Analysis and Representation” (in our system, the Sentence Segmention Module) involving basic information [1].

Our evaluation module is an evolution of the “Retrieval Functions & Ranking Module” that introduces the new concept of the sentence thesaurus and returns two output indexes: efficiency and quality as a “reality mining” approach applied to forensics data [9].

The system we are dealing with is also modeled taking into consideration Big Data, and its distributed structure can be adapted to a “map reduce” infrastructure [3].

## 7. Conclusions

What has been shown so far is an analysis based on Italian administrative judgments. This is because data on other types of sentences are unavailable. The described system (parser and sentence segmentator) is based on a knowledge base

in Italian, for this reason our modules do not currently work with other jurisdictions since the structure of judgments is different. Finally, the parser and the sentence segmentator should be adapted to the language and jurisdiction of reference. Then, the paper show how strong the need of a "predictable" jurisprudence is. The targets of knowledge, awareness and optimizations of rates and speeds can be reached through the use of ICT Tools [36]. Our work underlines that the efficiency and the quality of justice can be two important indexes to constantly monitor the improvement of legal processes. The approach has been implemented in a first software released as webapp that allows users to have an automatic opinion about keywords (sentence input) (the access to our system is available to research purpose), and providing a first public version is one of our most desired goals. In conclusion, the rapid development of information and communication technologies is offering new opportunities to improve the administration of justice with the availability of web services, the use of electronic ling, the electronic exchange of legal documents, the opportunity to consult the laws and the jurisprudence on-line: ICT can therefore be used to improve efficiency, access, timing and transparency.

## 8. Future developments

In the investigative and judicial fields, the analysis of correlations, the semantic enrichment and sentiment analysis provided with valuable tools for feedback of projective type: such a statistical analysis, thus, appears to be of fundamental support in the judicial activity.

The problems of data reliability, the provision of appropriate classifications in survey forms and, more generally, the quality of data are attributable, directly or indirectly, to the degree of computerization in statistical-judicial production.

Indeed, in the presence of a fully computerized detection system (also at the level of case management records) the possibility of transcription errors, manipulation and interpretation of the information required will drastically reduce.

On the other hand, the details related to the information collected could increase a result of a greater and more appropriate articulation of the detection patterns, and the activation of an automatic check on the consistency of the data would be possible even while the information is entered.

In Italy there have been massive investments in technological innovation with the promise of transforming justice in a quality service, increasing its ability to act in an effective, efficient, transparent way, and in line with the actual citizens' expectations, by allowing them to file a lawsuit and to forward it to the competent court, in a totally automated way.

This approach could give the chance to start new challenges related to Big Data analysis application fields and, in case of a possible future work, it is important to show and make Big Data scenarios rise considering the map/reduce approach with a real estimation of evolution of the system.

## Acknowledgments

**In loving memory of Professor Egidio Cascini**

*The farmer teaches that in life you should never prune the branch on which you usually place your staircase. Dear Professor Cascini, you will always be the beloved branch of our professional life, which always sustains our staircase and that we could never prune.*

*The authors would like to thank **Dr. Luca Pala**, for his support in the development of this work.*

## References

[1] O. Alhabashneh, R. Iqbal, F. Doctor, A. James, Fuzzy rule based profiling approach for enterprise information seeking and retrieval, Inf. Sci. 394–395 (2017) 18–37. July 2017
[2] R. Anacleto, L. Figueiredo, A. Almeida, P. Novais, A. Meireles, Step characterization using sensor information fusion and machine learning, Int. J. Interact Multimed. Artif. Intell. 3 (5) (2015) 53–60.
[3] A. Bechini, F. Marcelloni, A. Segatori, A mapreduce solution for associative classification of big data, Inf. Sci. 332 (2016) 33–55.
[4] V. Bevilacqua, V. Santarcangelo, A. Magarelli, A. Bianco, G. Mastronardi, E. Cascini, A semantic search framework for document retrievals (literature, art and history) based on thesaurus multiwordnet like, ICIC2011 (2011).
[5] F.J. Cabrerizo, R. Urena, J.A. Morente-Molinera, W. Pedrycz, F. Chiclana, E. Herrera-Viedma, A new selection process based on granular computing for group decision making problems, in: de Communications in Computer and Information Science, Springer, Andorra, 2015, pp. 13–24.
[6] C.E.C.f.t.E.o. Justice), European Judicial Systems: Efficiency and Quality of Justice, 2010 Data, Council of Europe Publishing, Strasbourg, 2012.
[7] C. Chang, Optimal information retrieval when queries are not random, Inf. Sci. 34 (December (3)) (1984) 199–223. 1984
[8] M.A. Chatti, Knowledge management: a personal knowledge network perspective, J. Knowl. Manag. 16 (5) (2012) 829–844.
[9] Y. Chen, N. Crespi, A.M. Ortiz, L. Shu, Reality mining: a prediction algorithm for disease dynamics based on mobile big data, Inf. Sci. 379 (2017) 82–93.
[10] T.M. Choi, H.K. Chan, X. Yue, Recent development in big data analytics for business operations and risk management, IEEE Trans. Cybern. 47 (2017) 81–92.
[11] T.-M. Choi, J. Gao, J.H. Lambert, C.-K. Ng, J. Wang (Eds.), Optimization and Control for Systems in the Big-data Era: Theory and Applications. International Series in Operations Research & Management Science, Springer, 2017.
[12] A. Constantinou, N. Fenton, M. Neil, Integrating expert knowledge with data in Bayesian networks: preserving data-driven expectations when the expert variables remain unobserved, Expert Syst. Appl. 56 (2016) 197–208.
[13] C. Cusatelli, The italian judicial offices productivity in almost 130 years of cognition civil procedures, J. Appl. Quant. Methods 6 (4) (2011).
[14] C. Cusatelli, M. Giacalone, Evaluation indices of the judicial system and ICT developments in civil procedure, Proc. Econ. Finance 17 (2014) 00885-5, doi:10.1016/S2212-5671(14)00885-5.
[15] G. Domeniconi, G. Moro, A. Pagliarani, R. Pasolini, Markov chain based method for in-domain and cross-domain sentiment classification, in: Proceedings of the 7th International Joint Conference, KDIR2015, 2015.
[16] R.M. Duwairi, Clustering semantically related classes in a heterogeneous multidatabase system, Inf. Sci. 4 June 2004 162 (2004) 193–210.

[17] M. Giacalone, S. Scippacercola, Big data: issues and an overview in some strategic sectors, J. Appl. Qual. Methods 11 (3) (2016) 1–17.
[18] M. Giacalone, A. Buondonno, A. Romano, V. Santarcangelo, Innovative Methods for the Development of a Notoriety System., in: A. Petrosino, V. Loia, W. Pedrycz (Eds.), Fuzzy Logic and Soft Computing Applications. WILF 2016, Lecture Notes in Computer Science, 10147, Springer, Cham, 2017.
[19] R.M. Gomes, A.P. Braga, H.E. Borges, Information storage and retrieval analysis of hierarchically coupled associative memories., Inf. Sci. 195 (15 July 2012) (2012) 175–189.
[20] M.T. Jones, Knowledge representation, Artificial Intelligence: A Systems Approach, Infinity Science Press, 2008.
[21] D. Jurafsky, Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Articial Intelligence), Second Edition, Prentice Hall, 2014.
[22] J. Kim, Scenarios in information seeking and information retrieval research: a methodological application and discussion, Lib. Inf. Sci. Res. 34 (2012) 300–307.
[23] O. Kolomiyets, M.F. Moens, A survey on question answering technology from an information retrieval perspective, Inf. Sci. 181 (24) (2011) 5412–5434.
[24] L. Laura, G. Me, Searching the web for illegal content: the anatomy of a semantic search engine, Soft Comput. 18 (1) (2015) 1–8.
[25] C. Lipizzi, L. Iandoli, J.E. Ramirez Marquez, Extracting and evaluating conversational patterns in social media: a socio-semantic analysis of customers' reactions to the launch of new products using twitter streams, Int. J. Inf. Manag. 35 (4) (2015) 490–503.
[26] B. Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, San Rafael, CA, USA, 2012.
[27] J.F. Lopez-Quintero, J.M. Cueva Lovelle, R. Gonzlez Crespo, et al., Soft comput (2016) a personal knowledge management metamodel based on semantic analysis and social information soft computing, (2016). 1–10, doi:10.1007/s00500-016-2437-y.
[28] J. Ma, G. Zhang, J. Lu, A state-based knowledge representation approach for information logical inconsistency detection in warning systems, Knowl. Based Syst. 23 (2) (2010) 125–131.
[29] M.C. Mochon, Social network analysis and big data tools applied to the systemic risk supervision, Int. J. Interact Multimed. Artif. Intell. 3 (6) (2016) 34–37.
[30] H. Moon, C. Lee, The mediating effect of knowledge-sharing processes on organizational cultural factors and knowledge management effectiveness, Perform Improv. Q. 26 (4) (2014) 25–52.
[31] J. Morente-Molinera, I. Perez, M. Urena, E. Herreda-Viedma, Creating knowledge databases for storing and sharing people knowledge automatically using group decision making and fuzzy ontologies, Inf. Sci. 328 (20) (2016) 418–434.
[32] OECD, OECD Compendium of Productivity Indicators 2017, OECD Publishing, Paris, 2017, doi:10.1787/pdtvy-2017-en.
[33] S.C. Pandey, G.C. Nandi, Convergence of knowledge, nature and computations: a review, Soft Comput. 20 (1) (2016) 319–342.
[34] Z. Saoud, S. Kechid, Integrating social profile to improve the source selection and the result merging process in distributed information retrieval, Inf. Sci. 336 (1 April 2016) (2016) 115–128.
[35] M. Schatten, Knowledge management in semantic social networks, Comput. Math. Organ. Theory 19 (4) (2013) 538–568.
[36] M. Sookhak, A. Gani, M.L. Khan, R. Buyya, Dynamic remote data auditing for securing big data storage in cloud computing, Inf. Sci. 380 (2017) 101–116.
[37] Y.-H. Tseng, Z.-P. Ho, K.-S. Yang, C. Chen, Mining term networks from text collections for crime investigation, Expert Syst. Appl. 39 (11) (2012) 10082–10090.
[38] B. Van Gils, H.A.E. Proper, P. van Bommel, T.h.P. van der Weide, On the quality of resources on the web: an information retrieval perspective, Inf. Sci. 177 (21) (2007) 4566–4597. 1 November 2007
[39] H. Wang, et al., Towards felicitous decision making: An overview on challenges and trends of big data, Inf. Sci. 367–368 (2016) 747–765. 1 November 2016
[40] M.-S. Wu, Modeling query-document dependencies with topic language models for information retrieval, Inf. Sci. 312 (10 August 2015) (2015) 1–12.