Unveiling Latent States Behind Social Indicators

Emanuele Di Buccio[1], Andrea Lorenzet[2], Massimo Melucci[1], and Federico Neresini[2]

[1] Department of Information Engineering, University of Padua, Italy
{dibuccio,melo}@dei.unipd.it
[2] Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua, Italy
andrea.lorenzet@gmail.com, federico.neresini@unipd.it

Abstract. The work reported in this paper aims at describing a project that leverages the potential of Information Retrieval and Machine Learning towards novel techniques that unveil the latent states of expert users such as sociologists and economists by means of indicators when the user is accessing large collection of newspapers, blogs, etc. An indicator measures the degree to which a certain latent state is present during interaction when exploring and searching an information repository. In this paper the state of a user is the particular condition that s/he is in at a specific context with reference to a problematic issue induced by the data s/he accessed to; risk is an example of state and a risk indicator aims at providing a measure of the degree to which articles examined by the user evoke risk in her/his mind. Observable attributes, e.g. keywords, click-through data or links, are the input data to model states and compute indicators. In this work, starting from some results of a software architecture designed to support sociologists in investigating Techno-scientific Issues in the Public Sphere, we will discuss some challenges and we will present a formal framework to address them where informative objects, e.g. news articles, states and attributes are uniformly modelled as vectors as it is customary in Information Retrieval or Machine Learning. This is our first step towards a long term objective, i.e. generalizing the well known Learning to Rank framework towards a Learning to Search framework which would encompass multiple and simultaneous states.

1 Introduction

Although Social Computing (SC) has a long history and dates back to, for example, Sadowski's work [14], the advent of big structured, semi-structured or unstructured multimedia data motivates, from the one hand, computer scientists to propose novel efficient and effective methods to monitor and analyse social, economic and political phenomena, thus developing a new research area called Social Computing. On the other hand, politicians, sociologists and economists are witnessing a major shift in social science research methodology thanks to vast, complex arrays of data to work with [1].

Not only the amount of data is rapidly increasing, the variety of types of data has also become larger and the degree of user interaction may be higher than

in the past. The typology of data that are accessed by social scientists, students or causal users such as World Wide Web (WWW) surfers includes structured data extracted from relational databases, semi-structured data received from eXtended Markup Language (XML)-encoded streams such as news feed, or unstructured data collected, indexed and delivered to end users by search engines as inter-linked pages. The different data management systems mentioned above can provide a variety of services to the end users who may search, browse and annotate text, images, video and music.

Together with users' behaviour data (e.g. click-through data or search session), these systems can be utilised to organize and analyse the users' thoughts and feelings about different controversial topics. For example, the vast and complex arrays of data such as social media can thus be utilized to create representations of techno-scientific controversies which are able to trigger intense public debates such as those concerning issues like climate change, Genetically Modified Organisms (GMOs), nuclear power.

When accessing to information by reading, browsing or annotating multimedia documents, the end users form a view or judgement about something, not necessarily based on fact or evidence, on the contrary, based on general feeling or opinion. These situations not only clearly blur the traditional boundaries among expertise, policy making, politics, and public opinion [10]. They also fuel an emotional *state* or reaction towards social issues and controversies. The knowledge of the user's state is not only interesting in itself, it is also crucial because it is such a state that often drives the user's decision about political and economic behaviours in different contexts (e.g. purchases, elections or party membership)[17].

In this paper, the state of a user is therefore the particular condition that s/he is in at a specific context with reference to a problematic issue documented by the data s/he accessed to. An example of user state is the condition that the user is in with reference to themes related to the "risk society". The risk society refers to the idea that within contemporary society Science and Technology (S&T) issues are imbued with fears and preoccupations about unforeseen effects, calling for a precautionary approach on the side of policy making, society, and the public [2]. It follows that, a user who is reading a newspaper editorial may be influenced by the editor's opinion on a topical issue, s/he may perceive risk, conflict, worry or controversy, and may collapse to the state in which s/he perceives some risk.

Note that understanding the user's state is different from the understanding the document's author opinion addressed by means of Sentiment and Opinion Mining and Analysis systems which aim to classify the opinion of a document's author who has implicitly expressed his opinion by means of a document. The user's state is not necessarily encoded in words, on the contrary, it might be understood by analysing different types of data ranging from click-through data to natural language queries in combination with document attributes. Moreover, our focus is not on users' mood that may be mined from short or tiny data such as tweets or "likes". Our interest is on readers who might not immediately comment and reveal their own feeling to the public, which is nevertheless of great interest

to *expert users* such as sociologists or economists who are asked to be acutely aware of the issues of our world.

This paper briefly reports on our advances in using SC techniques based on Information Retrieval (IR) and Machine Learning (ML) to support sociologists in investigating Techno-scientific Issues in the Public Sphere (TIPS). Starting from these initial results, it is our objective to further develop the framework and the TIPS software architecture to unveil the latent states that a user may experience when interacting with news articles that deal with techno-scientific controversies. In this context, IR may play a crucial role because it is naturally devoted to meet user informative needs stemming from the problem of understanding social phenomena, when these phenomena are encoded in large collections of inter-linked documents such as news, feeds, blogs, video and images.

Besides extending TIPS, we have got a long term vision. Quoting Liu [7], search engines by now go "beyond the pure relevance-based ranking of documents in their search results," since some prominent search engines "try to provide rich presentation of search result to users. When the ranked list is no longer the desired output, the learning-to-rank" – also known the application of ML to IR – "technologies need to be refined: the change of the output space will naturally lead to the change of the hypothesis space and the loss function, as well as the change of the learning theory. On the other hand, the new search scenario may be decomposed into several sub ranking tasks and many key components in learning to rank can still be used. This may become a promising future work for all the researchers currently working on learning to rank, which we would like to call learning to search rather than learning to rank." In this paper, we name this long term vision Learning To Search (LETS) where advanced systems are designed to support complex search task where the query-response paradigm is just a tactic within a search strategy. It is our opinion that considering multiple and simultaneous user' states is a step forward LETS and as a consequence an advance of SC technologies.

2 Motivations

For some years, a project called TIPS[3] has been carried out by us to develop, experiment, and implement automatic procedures for collecting, classifying, and analysing digital contents – mainly online news and user generated contents in social media – in order to monitor S&T issues and their evolution. TIPS collects online articles from six newspapers and classifies them according to their pertinence to S&T topics; currently the corpus is constituted of more than a million documents, collected since 2010. Fig. 1 reports an overview of the architecture components.

TIPS collects articles from online news such as RSS feeds associated to specific newspaper sections through a collector module. The articles are classified and a risk indicator is calculated to measure the degree to which the articles

[3] `http://hal.cloud.tilaa.com/tips/`

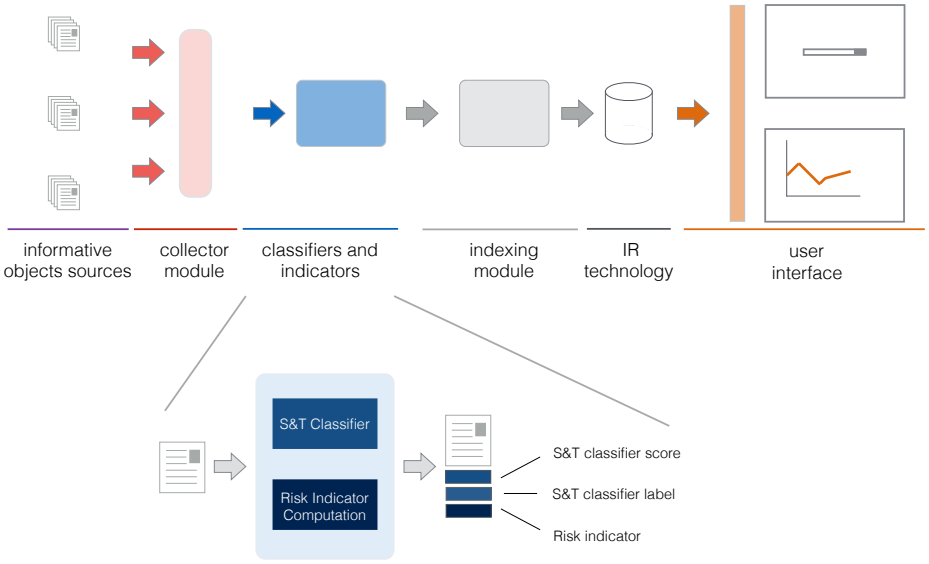


Fig. 1. Overview of the main TIPS architecture modules.

evoke risk in the users' mind. The articles are then enriched with the categories associated by the classifiers and the indicators. Furthermore, the articles are provided to an indexing module that stores the article attributes in a data repository that provides efficient access though meta-data and content-based search using IR technologies[4]. Additional layers can then be built on top of the IR module to provide meta-data and content-based access through a WWW user interface and to display dynamic charts representing indicator trends (Fig. 1).

The remainder of this section briefly describes how the risk indicator is currently instantiated in the current release of TIPS. A computer scientist may find the procedures implemented by this system rather rudimentary, however, the manual intervention is limited and it was required by the sociologists who wished to closely monitor the procedures as possible. A *risk indicator* relies on a set of keywords manually identified by the sociologists through the support of an unsupervised ML algorithm for the extraction of themes from an unstructured corpus. The keywords were selected by (i) retrieving documents through the query "risk", (ii) extracting topics and subtopics through an implementation of the HPAM topic modelling algorithm [9], and (iii) identifying terms in the conceptual area of risk by the manual inspection of the sub-topics. For example, a subset of the selected keywords could be: "infection", "danger", "alarm", "catastrophic". Each document is thus represented in terms of these keywords, as it is customary in many retrieval application domains.

[4] The current version of TIPS relies on ElasticSearch: `https://www.elastic.co`

An indicator described by the set of keywords $\mathcal{K}$ for the document set $\mathcal{D}$, will be computed as

$$\mathcal{I}_{\mathcal{K}}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \mathcal{I}_{\mathcal{K}}(d) \tag{1}$$

where

$$\mathcal{I}_{\mathcal{K}}(d) = \frac{1}{|\mathcal{K}|} \sum_{w \in \mathcal{K}} \frac{n_L(w,d)/B}{n_L(w,d)/B + K} \tag{2}$$

and $n_L(w,d)$ is the frequency of the term w in the document d; $n_L(w,d)$ is normalized by B: $(1-b) + b\frac{\mathrm{dl}(d)}{\mathrm{avgdl}(\mathcal{C})}$ where $\mathrm{dl}(d)$ is the length of the document d and $\mathrm{avgdl}(\mathcal{C})$ is the average document length in the corpus $\mathcal{C}$; $b \in [0,1]$ is a parameter that controls the weight assigned to the document length normalization. The $n_L(w,d)/B$ normalization has been introduced in [15]. The K's values control the effect of the term frequency on the indicator value for a document: the basic idea is to consider a non-linear dependence between the frequency and the indicator, thus delivering relatively high values already for small frequencies, and then "saturates" for large term frequencies. The indicator obtained for a document, $\mathcal{I}_{\mathcal{K}}(d)$, can be then used as additional document descriptor to perform meta-data based search.

In order to show a possible application of a risk indicator to actual news, in the remainder of this section we will report on the trends obtained for the risk indicator using an Support Vector Machine (SVM)-based classifier to identify S&T news articles. We considered a subset of articles collected by TIPS and published from January 1st, 2010 to December 31, 2015; the total number of articles constituting this subset is 630,549. A SVM-based classifier with linear kernel was trained on a sample of the news corpus that was manually labelled by a team of sociologists on the basis of their pertinence to S&T issues. The sample is constituted of 3817 documents; 1,393 were labelled as pertinent to S&T, while the remaining 2,424 as non pertinent. The effectiveness of the classifier was tested using a 60/40 split for training/test with 10 fold cross-validation. We then classified all the 630,549 articles using the learned classifier. The risk indicator was then computed for all the articles, for the subset of articles classified as pertinent and for those classified as not-pertinent. The evolution of the risk indicator on a monthly basis and in the time frame 2013-2015 is reported in Fig. 2; since the time granularity is one month, the document set $\mathcal{D}$ in Equation 1 for the lth time interval ti_l, denoted by $\mathcal{D}_{ti_l}$, is constituted by all the articles published in the lth month of the time frame 2013-2015; in the event of relevant and not-relevant trends, the document set is respectively constituted of the relevant and the not relevant documents published in the time interval ti_l.

One of the limitations due to using only content based descriptors is evident in Fig. 2: the trend of the risk indicator in the entire corpus differs from that in the subset of relevant articles, thus suggesting that the distribution of terms in risky documents can differ from that in relevant documents; therefore, frequency might be no longer the best evidence to capture the degree of risk.

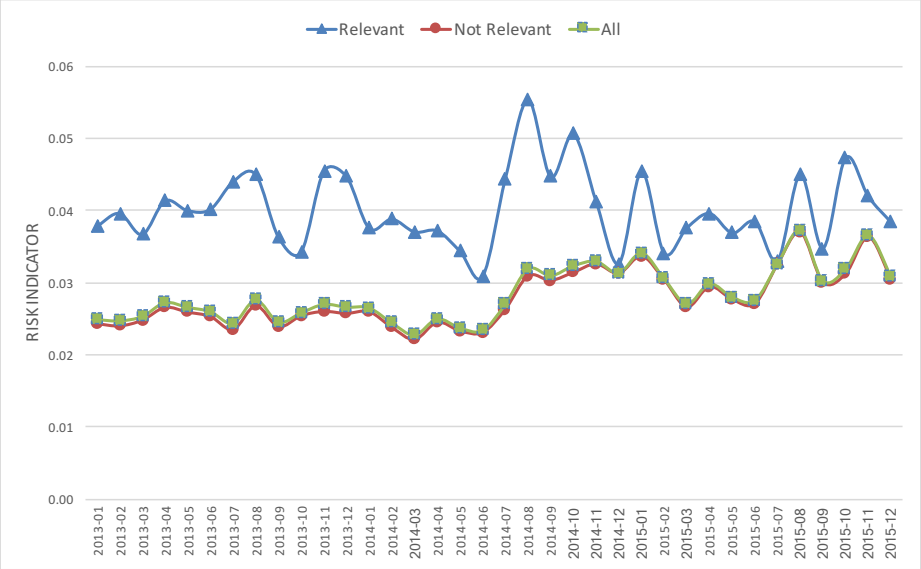


Fig. 2. Risk indicator in the time interval 2013-2015 for all the articles (green line), relevant (blue line) and not relevant (red line) to S&T issues.

Fig. 3 reports the variation of the risk indicator when computed for documents about "climate change" and documents about "nuclear power" for the subset of the news articles collected in TIPS and published from 2013 to 2015; also in this case the distribution of risky terms in the two subsets is different. However, even if the current indicator is able to capture a difference in terms of perception of risk in "climate change" and "nuclear power" related articles, the indicator is unable to explain why such difference exists. A more suitable representation of risk should be able to assists the users, e.g. specialists, also in this task. The framework introduced in the remainder of this paper aims to achieve such a goal through an explicit representation of *states*.

3 Challenges

TIPS has been the starting point of the contribution described in this paper. Since the early phases of the TIPS project, we realized that, beyond the sociologists' requirements, there is a great potential and some challenges of SC – mentioned in Section 1 – can be met if some current limitations of TIPS can be overcome. The main limitations of TIPS that are considered in this paper are the following ones:

Single state. Only one state (e.g. risk) can be at a time detected and measured from the information objects provided as input. The source software in principle is able to manage diverse states, but states are modelled as *independent* each

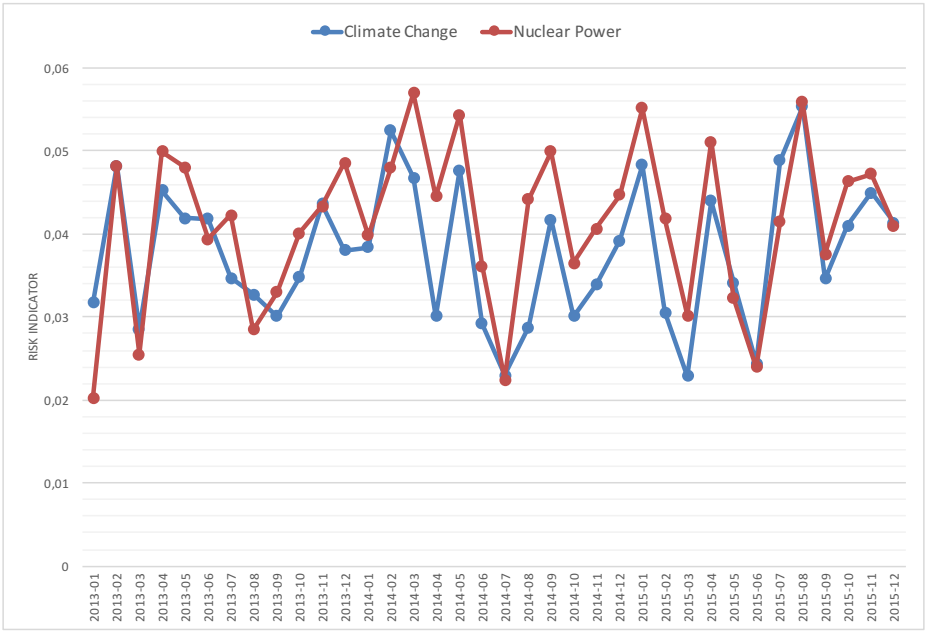


Fig. 3. Risk Indicator in the time interval 2013-2015 for articles about "climate change" and "nuclear power".

other. Moreover, some *manual intervention* is required to tailor the system to manage a specific state. The manual intervention to be made consists of compiling controlled keyword vocabularies such as the vocabulary of keywords evoking risk. Another state, e.g. "conflict" will require the compilation of another, distinct keyword list, and a group of experts in conflicts would be in charge of this compilation.

Content-based state modelling. The current implementation of TIPS only exploits content-based descriptors such as keywords, dates, and class labels. Because of the type of descriptors selected during system design, we have regarded risk as similar to relevance[5] and we used a normalized term frequency that worked well in general content-based IR [15] to compute the risk indicator.

However, in the light of detecting risk or other states, the current computation of indicators soon appeared rather simplistic. Normalized term frequency is not necessarily the best method to detect and measure risk and in general states other than relevance. For example, the distribution of terms in risky documents can differ from that in relevant documents and frequency might be no longer the best evidence to capture the degree of risk and actual interaction with the end user is ignored.

[5] A document is relevant to an information need when it carries information that help the user to meet his/her need

In contrast, the users' perception of risk may be better captured by the user behaviour during the interaction with the articles. Since an article might not include "risk" although it could evoke risk, or may include the word while not evoking the feeling, it is the complex of the users queries, click-through data, and other interaction features that may suggest the "riskiness" of the articles. Much information about the user's information need can be obtained through interaction [5,6]. Interaction can be, for instance, used in the actual computation of the indicator.

No simulation-based indicators. TIPS has been mainly designed for monitoring purposes by using IR and ML technologies. However, when studying the relationship between techno-scientific issues and the public's opinion, some tasks could benefit from the possibility to simulate specific scenarios, e.g. obtained by varying the degree to which news are "imbued" of risk.

In the current implementation, if the side effects of new scenarios were to be investigated, we should either build suitable synthetic documents or collect actual documents, and provide those documents as input to the pipeline constituted of the modules depicted in Fig. 1. However, some steps require manual intervention, which might be cumbersome or even impossible for specialists. For example, if a sociologist wanted to investigate how perception of risk will evolve if news about nuclear accidents were broadcasted, s/he should wait for actual news or generate synthetic news filled with keywords about nuclear accidents. Actual news documents would be difficult to assess because of the assessors' effort required to label a training set that is large enough to train TIPS. Although synthetic documents may in principle be generated about a topic that is traditionally perceived as source of risk, Natural Language Processing (NLP) technologies cannot be reliably utilized to generate these documents that a human expert can express a genuine perception of risk when s/he is asked to assess risk. If such technology existed, this challenge would be solved.

4 Contribution

In this paper, we address the challenges illustrated above and emerged from the TIPS project, i.e. the limitations that only one state can be investigated at a time, only informative content-based evidence is utilised to implement a state, and the impossibility of making prediction and simulation of what would happen to states when evidence will change. To the aim of facing these challenges, we introduce a vector-based formalism describing the main concepts of TIPS. The formalism is given in terms of definitions illustrated below.

Definition 1 (Information object). *An information object is any data container provided with an identifier.*

An information object is provided with an identifier to allow applications to connect the object, whereas search is based on the object's content. Examples of information objects are webpages, individual images, videos, music files, query

sessions, click-through data, and other user behaviour data. In this paper, an information object is symbolized by lower-case y and defined as a vector of the k-dimensional real space, since a component of y is a real number.

Definition 2 (Attribute). *An attribute is what we can directly observe from information objects such as documents or user interaction actions.*

Examples of attributes are informative content descriptors (e.g. keywords and terms), intra- and inter-object links, link anchors, meta-data, annotations, or tweets mentioning the information object (e.g. the article). The real components of a vector in $\mathbb{R}^d$ correspond to the attributes observed to measure an object. To obtain a complete formalism, in this paper, an attribute is symbolized by lower-case v and defined as a vector of the k-dimensional real space, since a component of v is a real number. For example, an attribute component may refer to a term frequency, a click occurrence, a colour code or an encoded sound fragment frequency. A set of independent attribute vectors form a vector basis and therefore any linear combination thereof is a vector of the same space. For example, the j-th basis vector of the k-dimensional space has 1 at component j and 0 elsewhere. An attribute may occur in an information object to a certain degree. Therefore, we have to introduce the following definition.

Definition 3 (Weight). *A weight is the degree to which an attribute is present in an informative object.*

A weight is thus a real number. In particular, we formalize an attribute as a vector of the k-dimensional real space. It is assumed an independence relationship between the attribute vectors, so that it is possible to formalize an information object as a linear combination of attribute vectors. If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are n attribute vectors, an information object vector can be written as

$$\mathbf{x} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n \tag{3}$$

where n is the number of attributes used to represent objects (e.g. the number of keywords) and the a's are the attribute vector weights measuring the degree to which an attribute describes an information object. Questions about a $\mathbf{v}$ can be answered when the vectorial representation $\mathbf{y}$ of a user is matched against $\mathbf{x}$, thus obtaining the corresponding a.

The main question is *how can we model states?* Contrary to attributes – which are manifest – *states* can only be indirectly – they are latent – observed from information objects.

Definition 4 (State). *A state is a latent characteristic of a user when interacting with an information object.*

The main thrust is that a state refers to both an information object and a user. As a state is a latent characteristic of an object-user pair, some attributes have to be observed to make the state explicit. However rich the description of information objects can be in terms of attributes, some hypotheses about the user's state when s/he is interacting with objects can be explained only if the

latent, unobserved states can be explicitly modelled, thus allowing to predict and simulate how states can evolve when the information objects – user interaction included – are observed. In this paper, a state is symbolized by lower-case z and defined as a vector of the k-dimensional real space, since a component of z is a real number. A state may occur in an information object to a certain degree. Therefore, we have to introduce the following definition.

Definition 5 (Indicator). *An indicator is the degree to which a state is present in an informative object.*

An indicator is thus a real number too. As objects, states and attributes are both placed in the same vector space, it is possible to represent an information object as a linear combination of m state vectors as follows:

$$\mathbf{x} = b_1\mathbf{z}_1 + \cdots + b_m\mathbf{z}_m \tag{4}$$

where m is the number of states, the $\mathbf{z}$'s are state vectors and the b's are indicators. Eq. (4) enables to compute each indicator using linear algebra operations. Thus, questions about a $\mathbf{z}$ can be answered when the vectorial representation $\mathbf{y}$ of a user is matched against $\mathbf{x}$, thus obtaining the corresponding b.

The formalism defined above stems from the theory of abstract vector spaces widely adopted in ML and IR and specifically in Learning to Rank (LETOR), which is based on vector-based attribute (also known as feature) spaces and discriminative learning [7]. We indeed leverage the potential of the combination of statistical learning and IR as implemented in LETOR since this combination has been proved to be efficient and effective. The definitions provided above means that attributes, objects and states will be represented in the same vector space. Using one single space allows us to obtain a uniform representation of attributes, states and objects by combining them using indicators and weights thereof, and to seamlessly apply operations on attributes, states and objects in a similar way as suggested by different retrieval and learning models.

Instead of depicting vectors using the usual arrows in a three-dimensional plot, we exploit the visual paradigm adopted in the current implementation of TIPS. Consider Fig. 4 which gives a pictorial description of the formalism mentioned above. First of all, there is a temporal axis along which states evolve as curves. Each state corresponds to a curve; "risk" corresponds to the red curve and "conflict" corresponds to the blue curve. The y-axis refers the indicator values; for example, the value of the indicator of conflict is 0.32 when the information objects are those observed on June 2014. The indicator value is depicted as a bullet placed on the curve. It is the result of a computational process that takes attribute weights as input.

Consider a scenario where a specialist user, e.g. a sociologist, working on the effect of S&T on the society and how the society affects the progress in S&T. Suppose that the sociologist's task is to study and comprehend the public perception of "climate change" in the last two decades. A possible sociologist's research question is: *how have the perception of risk related to "climate change" been changed in the last two years?* In order to carry out this investigation, a

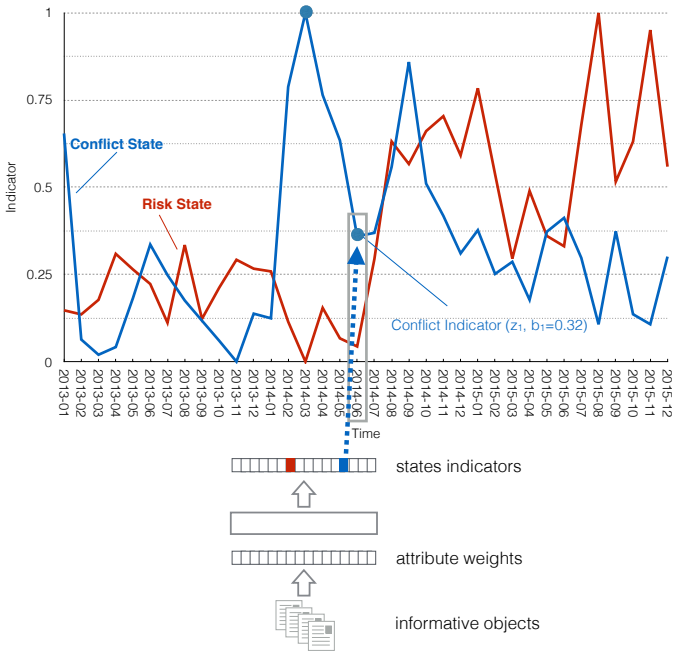


Fig. 4. The TIPS visual paradigm to describe the object-state-attribute formalism

possible source for information are articles published in newspapers; this is, for instance, the approach carried out in [11], where the authors investigated if it is possible to infer information about public opinion by looking at how the media discuss controversial technoscientific public issues. The task can be carried out by gathering all the articles published in the "most representative" newspapers and relevant to "climate change", manually analyse them and provide a qualitative discussion on the change, if any, on the perception of risk when considering this issue.

A parallel can be drawn between the notion of state and that of category of a classification system. However, state (e.g. risk) and category (e.g. nuclear energy) are different each other. Although the pertinence of a news to a category may be viewed latent, once the classification have been performed the fact that an article is pertinent to the category can be used as additional attribute to characterize the articles.

In contrast, a state, e.g. risk is that it is not directly observable from a document. Indeed, we are not considering articles about risk or on the notion of risk, but articles that are imbued or evoke risk. It is a latent state that can be evoked because of some words or combination of words occurring in the article that can trigger other issues or images related to risk, or because of two different viewpoints are presented with the underlying purpose of discrediting one of them.

The risk indicator is the degree to which the risk state is present in the user-object pair or in the interaction between the user and a set of documents.

Another difference between category and state is that a state lies in user interaction and it is not a static feature of an information object – a category may be viewed a static feature indeed – and does not change without changing the object content. Instead, state may change. A user who is reading news about nuclear energy may be or not be in a risk state depending on the personal or social context in which s/he interacting with information objects. The same apply to states other than risk such as conflict or economic crisis.

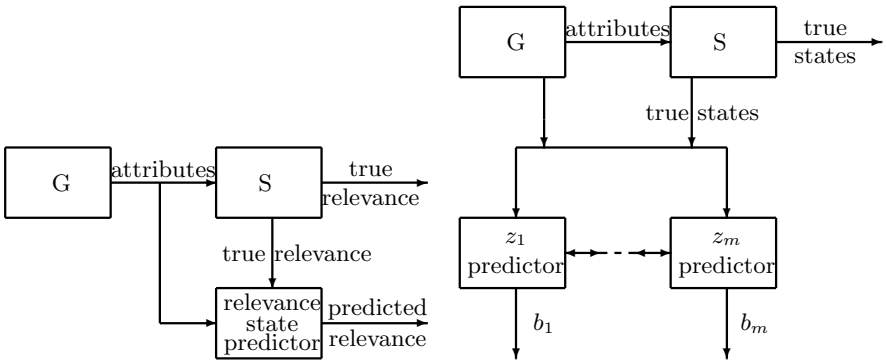
A similarity between states and categories (or classes) is multiplicity. Similarly to the simultaneous membership of an object to different classes, multiple states such as risk, conflict or economic crisis can be latent in the same object-user pair. Eq. (4) does indeed express the multiplicity and simultaneity of states in object-user pairs, where the z's can be adapted to the user's interaction. One of these z's may refer to relevance and the indicator b thereof may measure the degree to which x is relevant. Similarly, another z may refer to risk and the indicator b thereof may measure the degree to which x evokes risk. Thus, we have

$$\mathbf{x} = b_{\text{relevance}}\mathbf{z}_{\text{relevance}} + b_{\text{risk}}\mathbf{z}_{\text{risk}}$$

when an object x is represented in terms of latent states or

$$\begin{aligned}\mathbf{x} = {} & a_{\text{a term frequency}}\mathbf{v}_{\text{a term frequency}} \\ & + a_{\text{another term frequency}}\mathbf{v}_{\text{another term frequency}} \\ & + a_{\text{click frequency}}\mathbf{v}_{\text{click frequency}}\end{aligned}$$

The multiplicity of simultaneous states in an object requires a shift from the current state-of-the-art IR technologies based on LETOR to a novel paradigm called LETS. When only one state – relevance is the most important one in IR – is considered, ranking is the natural task that has to be automatized and LETOR is an appropriate approach to relevance-based document ranking, especially if applied to the WWW. The application to domains other than the WWW requires to model states other than and in parallel to relevance. Therefore, approaches other than LETOR may be useful if not necessary, since the users might no longer be casual users and the application domain might not be or only be about webpages to be ranked against relevance. In these contexts, ranked document lists may no longer be the most appropriate output as witnessed by the experience learned from TIPS. When the ranked list is no longer the desired output to answer one single state, LETOR needs to be refined and the change of the output space – multiple, simultaneous states – will naturally lead to the change of the hypothesis space – the space of functions – and of the loss function, as well as the change of the learning theory. The new search scenario may be decomposed into several sub ranking tasks – corresponding to the multiple simultaneous states although many key components in learning to rank can still be used. [7] Fig. 5 depicts some differences between the LETOR framework and the LETS framework. The former (Fig. 5(a)) considers one main



(a) LETOR framework (b) LETS framework

Fig. 5. The LETOR functional scheme (left) is based on a generator, G, of observable data (attributes), a supervisor, S, providing evidence about relevance, and a relevance predictor. The LETS functional scheme (right) integrates further states and indicators.

relevance predictor – there is one state, i.e. relevance – fueled by different attributes (or features). The LETS framework (Fig. 5(b)) instead includes more than one predictor, one predictor for each state. The prediction of these states may be simultaneous, thus requiring parallel optimizations and predictions.

5 Related Work

As the proposed project is interdisciplinary across sociology and computer science, contributions both in computational and social sciences are relevant to it. In [16] the "six degree of separation" phenomenon was confirmed at very large scale (Facebook Graph) and previously invisible social structures were captured. This example can be seen as inscribed in a broad process of developing new ways for the analysis of social phenomena built on the assumption that the Web is not another world - the virtual one - but is a constitutive element of social reality, and one increasingly relevant as argued in [12]. We use the expression SC to refer to the research activities that involve the analysis of social phenomena in digital resources [13]; another term is Computational Social Science.

Relevant contributions in modelling informative content were proposed in IR. In IR, Query Expansion (QE) techniques modify the initial query formulation by extracting from documents relevant (or assumed to be relevant) to the considered query additional descriptors to obtain a more effective information need representation. Many of these techniques are surveyed in [3] that points out that the adoption of QE on dynamic corpora is still an open issue.

Recent works focused on temporal Web dynamics and its application to IR, but they are mainly focused on Web user behaviour dynamics, on changing individual document content, or on the variation of the single term collection frequency over time. The TREC Knowledge-Based Acceleration track considers a

time ordered corpus; however, the task is filtering documents that would change the profile of people and organizations, and the list of entities is predefined - this project is not restricted to entity types.

Topic Models (TM) aim at automatically discovering the main "themes" in a document corpus. A well known TM is Latent Dirichlet Allocation (LDA) where documents are modelled as a distribution over a shared set of topics, which are themselves distributions over words generated by one of these topics. LDA assumes a fixed number of topics and the probability of seeing a topic is independent over time - in contrast, we will also address time. These issues are addressed in [4] but experiments are performed on relatively small datasets - in contrast, we will also address scalability.

This project focuses on techno-scientific controversies, i.e. issues able to trigger intense public debates, even in the case they address techno-scientific discussions such as those on climate change, GMOs, cloning, nuclear power merge and blur the traditional boundaries among expertise, policy making, politics, and public opinion [8]. Content analysis and the techniques traditionally used by social scientists to analyse textual corpora are the basis for developing novel indicators of techno-scientific controversies.

6 Final Remarks and Future Work

TIPS has been the starting point of an inter-disciplinary research project funded with the aim of providing expert users such as sociologists and economists with an effective and efficient system to investigate techno-scientific issues. Starting from this paper, we will pursue this objective and will also make a contribution at the level of methodology and system evaluation along two main directions.

We will address the problems related to the scenario simulation mentioned in Section 3 and will design, implement and evaluate methods for interacting with the representation of multiple and simultaneous states, e.g. risk and conflict. These methods will allow us to create scenarios by operating directly on state representations, thus avoiding the need to apply the entire pipeline for each scenario under investigation. To this end we will define a set of algebraic operators and implementation thereof within the functional scheme depicted in Fig. 5(b).

Besides explicitly considering interaction data when modelling states through ML algorithms, we will integrate interaction data to provide the user with information on the degree of a state present in the set of documents examined in the last sessions and how this degree differs from the "state distribution" in the overall corpus. In other words, interaction data can signal the tendency of a particular user to explore, say risky documents when performing a task or accomplishing a specific information goal. This signal can motivate the user to explore additional parts of the informative space in order to form his/her opinion on the issue and be "less subject" to the way the issue is presented in some venues — e.g. a particular set of newspapers or blogs.

References

1. R. M. Alvarez. Introduction. In *Computational Social Science*, pages 1–24. Cambridge University Press, 2016.
2. U. Beck. *Risikogesellschaft - Auf dem Weg in eine andere Moderne*. Suhrkamp, Frankfurt/Main, 1986.
3. C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1–50, Jan. 2012.
4. A. Dubey, A. Hefny, S. Williamson, and E. P. Xing. A nonparametric mixture model for topic modeling over time. In *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA.*, pages 530–538, 2013.
5. D. Kelly and J. Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, 2003.
6. M. Lalmas and I. Ruthven. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(1):95–145, 2003.
7. T.-Y. Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.
8. A. Lorenzet. Fear of being irrelevant? Science communication and nanotechnology as an internal controversy. *Journal of Science Communication*, 11(4), 2012. C04.
9. D. Mimno, W. Li, and A. McCallum. Mixtures of Hierarchical Topics with Pachinko Allocation. In *Proceedings of ICML '07*, pages 633–640, 2007.
10. F. Neresini. And man descended from the sheep: the public debate on cloning in the italian press. *Public Understanding of Science*, 9(4):359–382, 2000.
11. F. Neresini and A. Lorenzet. Can media monitoring be a proxy for public opinion about technoscientific controversies? The case of the Italian public debate on nuclear power. *Public Understanding of Science*, 25(2):171–185, 2016.
12. R. Rogers. *The End of the Virtual*. Amsterdam University Press, 2009.
13. R. Rogers. *Digital Methods*. MIT Press, 2013.
14. G. Sadowsky. Future developments in social science computing. In *Proceedings of Spring Joint Computer Conference*, pages 875–883, 1972.
15. A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29, 1996.
16. J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.
17. C. Warshaw. The application of big data in surveys to the study of elections, pubic opinion, and representation. In *Computational Social Science*, chapter 1, pages 27–50. Cambridge University Press, 2016.