# Detection of Cognitive Features from Web Resources in Support of Cultural Modeling and Analysis

Antonio Penta, Nigel Shadbolt, Paul Smart
School of Electronics and Computer Science
University of Southampton,
Southampton SO17 1BJ, UK
{ap7,nrs,ps02v}@ecs.soton.ac.uk

Winston R. Sieck
Applied Research Associates, Inc.
1750 Commerce Center Blvd
Fairborn, Ohio, 45324, USA
wsieck@ara.com

## ABSTRACT

The World Wide Web serves as a valuable source of culture-relevant information, which can be used to support cultural modeling and analysis activities. Part of the challenge in exploiting the Web as a source of culture-relevant information relates to the need to detect and extract information about beliefs, attitudes, and values from a variety of different resources. The Web, thus, features a rich variety of information resources, and these are seldom categorized with respect to the dimensions in which cultural analysts are interested. Exploiting the Web as a source of culture-relevant information therefore requires techniques and approaches that enable cultural analysts to extract relevant information and organize extracted content in various ways. In this paper, we outline an approach to assist cultural analysts in the extraction and organization of relevant information. We show techniques that can be used to extract information of the attitudes, beliefs, and values of individuals, and how this data can, in turn, be used to support cultural modeling and analysis.

## Categories and Subject Descriptors

H [**Information Systems**]: *Social Computing, Cultural Modelling, Cognitive Features Detection*; H.3 [**Information Search and Retrieval** ]

## 1. INTRODUCTION

The World Wide Web (WWW) serves as a valuable source of culture-relevant information, which can be used to support a number of cultural modeling and analysis activities. A number of factors, however, militate against the widespread use of the Web in cultural analysis contexts. One difficulty relates to the fact that Web content is seldom represented and organized in ways that support cultural modeling and analysis. If cultural analysts therefore wish to test specific hypotheses regarding the distribution of beliefs, values and attitudes (what we collectively refer to as "*Cognitive Features*") among different groups, they are often prevented from doing so in a Web context because the data is simply not available in the right format. Typically, culture-relevant information is embedded in resources containing other kinds of content, and this makes systematic forms of data analysis highly problematic. Ideally, what is required are representational schemes that enable cultural analysts to flexibly manipulate data in ways that support hypothesis testing and theory development. A second, not altogether unrelated concern, associated with the use of the Web as a source of culture-relevant information relates to the fact that relevant data is often not explicitly represented in the target resources. For example, if we are looking for evidence of particular Cognitive Features in natural language resources, then we will often have to analyze the meaning of the source text; seldom will target Cognitive Features be represented in such a way that they can be easily detected by automated processing techniques. In light of these difficulties, it is important to develop a range of information extraction, representation and manipulation capabilities. Such capabilities need to be flexible enough to extract a range of Cognitive Features, and they need to be sensitive enough to detect those features even when the target features are "hidden" in natural language texts (a problem that is akin to the detection of weak signals in a lot of background noise). Finally, information manipulation capabilities are required to support hypothesis testing and cultural modeling activities. The development of these capabilities will support the use of the Web as a resource for cultural analysis and cultural model development. In this context, the aims of the paper are: i) to propose a general framework that can be used to support the detection, extraction and representation of Cognitive Features; ii) to show how we can use statistical techniques to implement the proposed framework; iii) to show in a preliminary case study how the Cognitive Features can be useful patterns to detect members belonging to an extreme religious domain. To the best of our knowledge, ours is the first attempt to propose a computational approach to address the cognitive models by a cultural framework. An interesting survey of cultural influence in social behaviour is presented in [9]. Obviously, there is a rich literature concerning the extraction of particular bodies of information from Web-based sources [1]; however, most of the techniques that are described in the information extraction literature focus their attention on extraction algorithms while ignoring the specification and selection of features that can be used to support extraction goals. In this paper, we use an approach that is based on the notion of Topic Models. Blei et al [2] first used this approach to represent a document as a mixture


*MEDES'11* November 21-24, 2011, San Francisco, USA

of topic distributions, a topic being a statistical distribution over the words belonging to the vocabulary of a considered corpus. A number of variants of this approach have been proposed in the literature; for example [6]. In the current paper, we adopt the notion of Topic Models, but we extend the notion to include a new graphical model component, and we also give a specific meaning to the distributions used in our models (this is something that is rarely discussed in the context of Topic Model research).

The current paper is organized as follows: in Sections 2 and 3 we provide an overview of our approach to Web-based knowledge extraction in support of cultural modeling and analysis; in Sections 4, 5, and 6 we present the methodology used to represent the analytic substrate of the information extraction process for the Web-based textual sources; in Section 7 we describe the technical approach used to analyze the text sources; and in Section 8 we present a specific example of our approach focused on the domain of religious extremism.

## 2. CULTURAL ANALYSIS AND COGNITIVE FEATURES

We adopt an epidemiological approach to culture, which sees inter-individual similarities in cognition as the basis for cultural groupings [8]. A fundamental assumption of this perspective is that shared developmental experiences lead to important similarities in the mental representations (e.g. concepts, beliefs and values) that are distributed among members of a population. The Web appears as the right place to study how ideas are spread among behavioral norms, discussions, interpretations, and affective reactions within specific populations. We are interested in models that are able to elicit, analyse, and represent the beliefs, values, and cognitive concepts that are shared by members of a cultural group and how these affect their decisions or how they are connected. First, let us give an informal definition of what is, for us, a cultural group:

DEFINITION 1. *A Cultural Group is a collection of people who are grouped together by virtue of their similarity along specific cognitive dimensions; e.g. commonality of beliefs, attitudes and values.*

Now let us describe how we model the relationship among a cultural group and the cognitive signatures of its individuals in our perspective. We start from the modelling approach that was developed in [8]. The approach is called Cultural Network Analysis (CNA). In CNA, a conceptual model based on belief network is used to show the cultural knowledge within a population. We model the Cognitive Features as follow:

DEFINITION 2. *A Cognitive Feature (CF) is one of the following structures: 1) a triple $\langle B, v, \delta \rangle$, with $B$ being a belief, $v$ one of the values of $B$ and $\delta \in \mathbb{R} \cap [-1, +1]$ a measure of the value or belief perception in a group or individual: negative ($-1 \leq \delta < 0$), positive ($0 < \delta \leq 1$) or neutral ($\delta = 0$); ii) a triple $\langle C, E, p \rangle$, with $C$ and $E$ being cause and effect, respectively, of the casual relationship $C \rightarrow E$ and $p \in \{+, -\}$ is a negative ($-$) or positive ($+$) polarity.*

According to our epidemiological approach we simply define our model as follows:

DEFINITION 3. *A Cultural Model ($\mathcal{M}$) is a set of Cognitive Features*

For example, to understand what is meant by the terms "belief" and "value", let us consider the religious domain. In this domain, we introduce some beliefs knows as *meta-cognitive beliefs*. The terms *meta-cognition* refers to the beliefs about how one thinks and learns [7]. In particular, these beliefs are the ones that affect the cognitive processes that govern feelings of confidence in world-views. In Table 1 are reported the meta-cognitive beliefs that we introduce together with their values. We reported just an example of the meaning of these beliefs such as the one related to the belief *Knowledge*. Knowledge belief has two values: i) *Maintenance* that represents ideas that emphasize the priority and continuance of long-established conceptions of the world used to block new information, interpretations ; ii) *Change* that represents a belief that emphasises knowledge acquisition and change at the individual and cultural levels and it implies that existing beliefs may be wrong and incomplete, or no longer fit with current situations. Further explanation of those metacognitive beliefs can be found in [7]. Examples of causal relationship in a cultural environment can be the triple $\langle$*Religion*, *Innovation*, -$\rangle$. This means that we can have a decrease in the Innovation proportional to a rise of a Religion. For example, if we process the information coming from two kind of cultural groups characterized by an extremist ($G_E$) or moderate ($G_M$) vision about the meaning of the religion in the world, we can imagine to obtain the following cultural models $\mathcal{M}_E$={$\langle$*Knowledge*, *maintenance*, 0.9$\rangle$,$\langle$ *Judgement*, *authority*, 1$\rangle$, $\langle$*Religion*, *Innovation*, -$\rangle$,$\langle$*War*, *Honour*, +$\rangle$} for $G_E$.
$\mathcal{M}_M$={$\langle$*Knowledge*, *change*,1$\rangle$,$\langle$*Coherence*, *Diversity*, 0.6$\rangle$, $\langle$*Thinking*, *Freedom*, +$\rangle$,$\langle$*Democracy*, *Religion*, +$\rangle$} for $G_M$.
s, without taking into account the idea of their positive, negative or neutral attitude.

| Meta-Cognitive Beliefs | | |
|---|---|---|
| | Knowledge | Coherence |
| Values | Maintenance (KM)<br>Change (KC) | Homogeny (CH)<br>Diversity (CD) |
| | Information Exchange | Judgement |
| Values | Separation (IES)<br>Interaction (IEI) | Authority (JA).<br>Independence (JI) |

**Table 1: The meta-cognitive beliefs and their values used to categorized the cultural signals in the religious domain.**

## 3. COGNITIVE FEATURE DETECTION

One way in which Cognitive Features are manifested on the Web is in response to the occurrence of particular events, for example, military interventions, terrorist attacks, public protests and so on. These events elicit responses that reveal something about the beliefs, attitudes and values of the respondents to the event in question, and they therefore reveal something about respondent cognitions. Given the aforementioned cognitive characterization of culture, we can see that individuals' responses to particular events can be a valuable source of cultural information. This is one reason

why the Web serves as a source of culture information. The advent of Web 2.0 has supported greater participatory interaction with the Web, and enabled individuals to contribute to Web content. If information extraction technologies can be used to extract information about individual cognitions from the kind of resources in which individuals typically express their views (for example, blogs, twitter feeds, discussion forums, and so on), then we may be able to detect some of the features that are important for cultural analysis and modeling. In order to support this detection process, we are interested to process sources that deliver signals that we define as *cognitive*. Let us first introduce informally what we mean for cognitive signals as follows:

DEFINITION 4. *The Cognitive Signals are all the messages where the people elicit their thinking referable to a cultural knowledge within a population. These messages have to be automatically processed and can be exchanged using different media.*

Examples of sources, that convey Cognitive Signals and how these can be used to detect the relationships between the individuals and a cultural group, are presented in Figure 1. For example, an image can be an important indicator of the relationships between Web page authors and the cultural groups to which they belong. In particular, a more complex analysis is required based, for example, on how much these signs are used among linked members or for example in which position they are depicted. Intuitively, an image on the title banner is more valuable than others. Example of images are: i) logos related to political organizations (Figures: 1.a, 1.b); ii) flags (Figures 1.c, or of a 1.d); iii) symbols of terrorist groups (Figures 1.a,1.b); vi) images of historical characters (Figure 1.h). Another example, in this case in video format, is represented in Figure 1.g. It is the famous speech about freedom in the Braveheart movie. A high degree of content sharing among a community provides an indication of how important notions such as freedom are among a group and how they think about freedom. Most of these signals can not be processed separately, so a multimodal analysis using different media is required. Processing different signals over different media channels can provide important inputs to cultural modeling and analysis. In spite of the importance of multi-modal analysis, much of the input for Cognitive Feature detection will probably come from text-based sources. In this paper, we focus our attention on the Cognitive Features extracted from signals related to text sources. Examples of text sources that are relevant from the cultural point of view are depicted in Figure 1.i and Figure 1.l. These sentences reveal the views of content authors that reflect their membership of particular cultural groups (for example, moderate or extremist religious groups). In this setting, the Cognitive Features detection process aims to model, extract, and process those Cognitive Signals in order to detect the Cognitive Features and eventually structure all the results of this process in what we call *Cognitive Patterns*. Let us give a formal definition of this object.

DEFINITION 5. *Let us consider a Cultural Model $\mathcal{M}$ and one of its Cognitive Feature $\tau$. A Cognitive Pattern ($\mathcal{P}^{\tau}_{\mathcal{M}}$) associated to $\tau$ belonging to $\mathcal{M}$ is a set of triples as $\langle r, \tau, \mu \rangle$, where $r$ a source containing a Cognitive Signals referable to an individual or group within a population and $\mu \in [0,1] \cap \mathbb{R}$ a measure of how $r$ is reliable to be a representative of $\tau$ on the considered individual, group or population.*

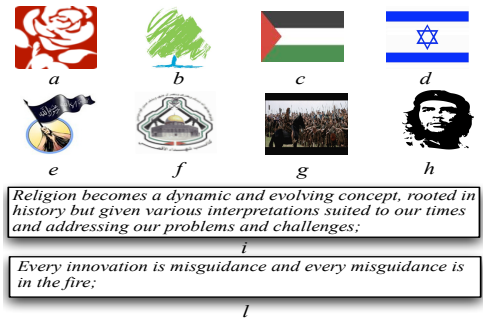


**Figure 1: Examples of Cognitive Signals in different media formats.**

We note that in this setting the resource $r$ can be any data belonging to a group in any format that an expert identifies as a valuable source of cultural information. Then, this detection process has the aim to populate a "Cultural Pattern Database" ($\mathcal{P}db$) where all this knowledge is stored and updated by domain experts. Now, let us describe how we deal with the Cognitive Signals related to the text sources.

# 4. THE TEXT WEB SOURCES

In this section we explain how we model and extract signals from the text related to a web page. First we give a more formal definition of how we model the text messages and then how we extract our model from a text document.

## 4.1 Text Signal Modeling

We model the text using a linguistic model known as the N-gram approach [5]. In this model, the text is divided into structures, known as *gram elements*, which are formed by tokens extracted from the text. Let us consider a text fragment and assume that we extract from it some gram elements, looking at the words as linguistic tokens. Firstly, we use the term *gram elements types* to indicate the type of ngram extracted. A gram element type can be associated to one of the following *category*: uni-gram, bi-gram or tri-gram. Let us introduce the definition of *Text Signal* as follows:

DEFINITION 6. *A Text Signal is a set of gram elements belonging to the same category. In particular, we use the following symbols: i) $\mathcal{T}$ for the Text Signal made by uni-gram; ii) $\mathcal{T}^2$ for the Text Signal made by bi-gram; iii) $\mathcal{T}^3$ for the Text Signal made by tri-gram.*

Now, looking at the Part of Speech (PoS) tag [5] associated with each word belonging to a Text Signal, we can differentiate them as follows:

DEFINITION 7. *Let us consider the following elements $(w_i) \in \mathcal{T}$, $(w_i,w_j) \in \mathcal{T}^2$, $(w_i,w_k,w_j) \in \mathcal{T}^3$, and let us attach to all of them their Part of Speech (PoS) labels as follows: : $(w_i/l^1_i)$, $(w_i/l^2_i,\ w_j/l^2_j)$ and $(w_i/l^3_i,w_k/l^3_k,w_j/l^3_j)$. We can differentiate these Text Signals using the computed PoS labels as follows:*

- *A Text Entity Signal is a subset of those elements belonging to $\mathcal{T}$ or $\mathcal{T}^2$ or $\mathcal{T}^3$ that fulfil the following conditions: i) for $\mathcal{T}$, $l^1_i$ is a noun or proper noun; ii) for $\mathcal{T}^2$, we have that both $l^2_i$ and $l^2_j$ are nouns or proper*

*nouns; iii) for $\mathcal{T}^3$, we have that both $l_i^3$, $l_j^3$ are nouns or proper nouns and $l_k^3$ is a verb. We use the following symbols $\mathcal{E} \subseteq \mathcal{T}$, $\mathcal{E}^2 \subseteq \mathcal{T}^2$, $\mathcal{E}^3 \subseteq \mathcal{T}^3$ to refer to those subsets.*

- *A Text Sentiment Signal is a subset of those elements belonging to $\mathcal{T}^2$ or $\mathcal{T}^3$ that fulfil the following conditions: i) for $\mathcal{T}^2$, we have that $l_i^2$ is a noun or proper noun and $l_j^2$ is an adjective; ii) for $\mathcal{T}^3$, we have or $l_i^3$ is a noun or proper noun, $l_k^3$ is a verb and $l_j^3$ is an adjective, or $l_i^3$ is a noun or proper noun, $l_j^3$ is an adjective and $l_k^3$ is adverb, or $l_i^3$ is a noun or proper noun and both $l_j^3$ and $l_k^3$ are adjectives. We use the following symbols $\mathcal{S} \subseteq \mathcal{T}$, $\mathcal{S}^2 \subseteq \mathcal{T}^2$, $\mathcal{S}^3 \subseteq \mathcal{T}^3$ to refer to those subsets.*

For example, using the sentence in Figure 1.i a Text Entity Signal can be {*religion*, *history*, (*religion, interpretation*), (*religion, root, history*)} and a Text Sentiment Signal can be {(*religion, dynamic*), (*religion, become, dynamic*)}.

## 4.2 Text Signal Extraction

We can now describe the process to extract Text Signals from an unstructured text document $d$. We first pre-process $d$ by sending it to a standard Natural Language Processing (NLP) pipeline made up of the following components: Sentence Tokenizer, Word Tokenizer, Part of Speech tagger, Stop Word Eliminator, etc. (more details about these NLP steps can be found in [5]). [1]. After these phases, we represent each of the sentences of $d$ as a vector of words with a related vector corresponding to the PoS annotation of each word. For example, for the sentence $s_i$ in $d$, we have a vector $\overrightarrow{x}_i$ made by the words plus an associated vector $\overrightarrow{x}_{Li}$ of the same cardinality of $\overrightarrow{x}_i$ such that $\overrightarrow{x}_{Li}[k]=l_k$ is the PoS label of the word $w_k=\overrightarrow{x}_i[k]$. The elements of $\overrightarrow{x}_i$ are the words filtered in the previous NLP pipeline. Then, we can derive, for each vector, a Text Signal. In the case of $\mathcal{T}$, we have just the elements of a vector $\overrightarrow{x}$, instead for $\mathcal{T}^2$ and $\mathcal{T}^3$ we reduce the possible number of binary and ternary combinations of elements belonging to a vector, by choosing a maximum linguistic dependency that have to be considered among its words. In particular for $\mathcal{T}^2$ and $\mathcal{T}^3$ the strategy used to extract bi-grams and tri-grams from a vector is depicted in Figure 2. In particular, we choose a *dependency window* $w$ and then from this value we can compute the indexes used to extract the bi-grams and tri-grams. In Figure 2, the indexes for a generic step $i$ of our extraction process are depicted both for bi-grams and tri-grams. Note that in Figure 2 we choose the same dependency window $w$ for both bi-grams and trigrams. In general , if we have a vector of cardinality $N$, we extract: i) $(N-w_b)w_b+\frac{w_b(w_b-1)}{2} \ll \binom{N}{2}$ number of bi-grams if $N > w_b$ being $w_b$ the dependency window for a bi-gram, otherwise all the different combinations; ii) $(N-2w_t)w_t^2+w_t^2(w_t-1) \ll \binom{N}{3}$ number of tri-grams if $N > 2w_t$ being $w_t$ the dependency window for a tri-gram, otherwise all the different combinations. After computing $\mathcal{T}$, $\mathcal{T}^2$, $\mathcal{T}^3$ using a vector $\overrightarrow{x}$, we can apply a filter based on the information computed in $\overrightarrow{x}_{Li}$, in order to obtain the Text Entity Signal and Text Sentiment Signal as described in Definition 7.

[1]If there is a "not" that comes before an adjective, we collapse the negation with the adjective to create unique words, for example "not good" becomes "not_good". This is important in order to deal with the negation of an adjective, which can completely change the meaning of the adjective itself

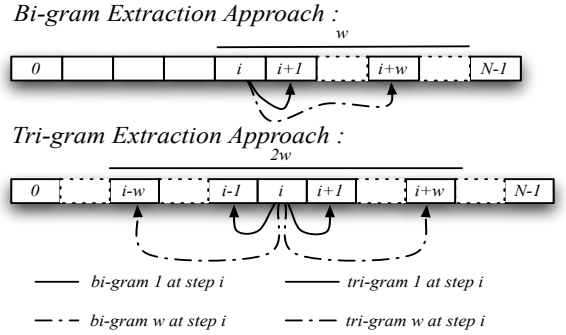


**Figure 2: Approach to the extraction of bi-gram and tri-gram from a vector of words.**

# 5. COGNITIVE ANNOTATION

The problem now is to understand how we can use the previous text signals in order to detect our Cognitive Features. We use a supervised approach, where the cultural analysts give an initial subset of annotated resources that are used by the methods described in the next sections. We call this initial set the *Cognitive Annotations*. For example a cultural expert can initially select from the Web a text fragment, such as the one belonging to a blog, because this is valuable to describe the Cognitive Features $\tau_j$ within the Cultural Model $\mathcal{M}_i$. This annotation can be initially given by the analyst in a way similar to how we describe the Cognitive Patterns. For example using the Cultural Model defined previously $\mathcal{M}_E$ and $\mathcal{M}_M$, their annotation can be structured as follows: $\mathcal{P}_{\mathcal{M}_E}$={⟨"*religion is distorted*", ⟨Knowledge, maintenance, 0.3⟩, 1⟩}; $\mathcal{P}_{\mathcal{M}_M}$= {⟨"*knowledge is the first step in firm belief and conviction*", ⟨Knowledge, change, 0.8⟩, 0.6⟩}. We consider a real scenario in which the resources are unstructured texts and the cultural analysts can be different so we need to define how we process these annotations in order to build particular valuable patterns that are related with these annotations. Let us describe how we process a text annotation $tf_k$ that a cultural analyst made for the Cognitive Features $\tau_j$ belonging to a Cultural Model $\mathcal{M}_i$ and how we define these Cognitive Annotations. In particular we suppose that each text-based resource used in the annotation is a sentence. We process this unstructured knowledge to extract the Entity and Sentiment Text Signals in the same way described in Section 4 for the Text Signals, using the NLP pipe, the bi-grams/tri-gram extraction process and the PoS filter. After this process, we have the following sets $\mathcal{T}$, $\mathcal{T}^2$ and $\mathcal{T}^3$ or $\mathcal{E}$, $\mathcal{E}^2$, $\mathcal{E}^3$, $\mathcal{S}^2$ and $\mathcal{S}^3$. Now we build a Cognitive Annotation as a special Cognitive Pattern $\mathcal{P}_{\mathcal{M}_i}^{\tau_j}$, where a triple is $\langle \mathcal{A}_i^k, \tau_j, \mu_i^k \rangle$. Being $\mathcal{A}_i^k$ a set of gram elements computed over the initial resources $tf_k$ and $\mu_i^k$ a new reliable measure. In particular we divide these annotations into i) *Simple Text Cognitive Annotation* if $\mathcal{A}_i^k$ is one of the following sets: $\mathcal{T}$, $\mathcal{T}^2$, $\mathcal{T}^3$ ; i) *Entity Text Cognitive Annotation* if $\mathcal{A}_i^k$ is one of the following sets: $\mathcal{E}$, $\mathcal{E}^2$, $\mathcal{E}^3$; iii) *Sentiment Text Cognitive Annotation* if $\mathcal{A}_i^k$ is one of the following sets: $\mathcal{S}^2$, $\mathcal{S}^3$. Let us now explain how we compute the new reliable

measure $\mu_i^k$. In order to compute this new measure we use the previous annotations. The value $\mu_i^k$ in $\langle A_i^k, \tau_j, \mu_i^k \rangle$ can be computed as follows:

$$\mu_i^k = \alpha_1(avg(NMI_{A_i^k})) + \alpha_2(\mu_k) \quad (1)$$

In this equation, we use a convex combination, $\alpha_1 + \alpha_2 = 1$, of the average ($avg$) value of the Normalized Mutual Information [5] computed for each gram element in $A_i^k$, i.e $NMI_{A_i^k}$. With $\mu_k$ we mean the initial value associated by a domain expert to annotate the resource $tf_k$. We compute the Normalized Mutual Information as follows: we indicate with $g$ a gram element belonging to $A_i^k$ and with $\tau_j$ the related Cognitive Features. We define for $g$ and $\tau_j$ two binary random variables $X_g$ and $Y_{\tau_j}$ respectively and then we compute an associated contingency table, such as the one depicted in Figure 3. In this table, we represent the frequencies related to how much a gram $g$ is used to describe $\tau_j$ or not.[2]. In particular, the sub-references of 0 and 1 used in the table in Figure 3 are used to indicate the absence or presence of our variables. For example, if we would measure the events

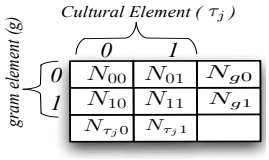


**Figure 3: Contingency Table used to compute the Normalized Mutual Information.**

in which the considered gram element is used to describe $\tau_j$, the joint probability is $P(X_g = 1, Y_{\tau_j} = 1) = \frac{N_{11}}{N_{tot}}$, being $N_{11}$, the number of times the gram element $g$ is used in the resources associated with the Cognitive Features $\tau_j$ and $N_{tot}$ the total number of gram elements of the same gram element type of $g$ stored in all the annotations. Then, we compute the Normalized Mutual Information for $g$ and $\tau_j$ as follows:

$$NMI(X_g, Y_{\tau_j}) = \frac{MI(X_g, Y_{\tau_j})}{min(H(X_g), H(Y_{\tau_j}))} \quad (2)$$

$$MI(X_g, Y_{\tau_j}) = \sum_{r \in \{0,1\}} \sum_{l \in \{0,1\}} P_{12} log(\frac{P_{12}}{P_1 P_2})$$

where $H(*)$ is the entropy of a random variable $\{*\}$, $P_{12} = P(X_g{=}r, Y_{\tau_j}{=}l)$, $P_1 = P(X_g{=}r)$ and $P_2 = P(Y_{\tau_j}{=}l)$ and $MI$ the Mutual Information. We note that this procedure is useful to understand what are the best annotations that can be used within our process. This procedure can also be triggered every time we have a new annotation in order to use the best knowledge collected by the domain expert. We call "Cognitive Annotation Database" ($\mathcal{CA}db$) the place where all this knowledge is stored and updated by domain experts.

# 6. THE GRAM ELEMENTS DISTANCES

In this section, we explain how we compare the gram elements, such as the ones belonging to the sets introduced in the above sections. Computing semantic distances among the words in the extracted gram elements can require a lot of time due the complexity of the measures related to the navigation of the knowledge used to support this computation. In order to optimize this step, we used a hybrid approach based on a linguistic and semantic distance. This approach has the aim to choose a different distance computation according to the PoS labels associated with each word of our gram elements. We define the following strategies:

- *Strategy 1 (S1)* based on the *Edit similarity* that measures how many linguist operations we need to use in order to transform one word into another one.
- *Strategy 2 (S2)* based on the *Jaccard similarity* that measures how many elements two sets have in common. In particular, for each word we build a set with the synsets retrieved from the WordNet [4] database that are connected at maximum distance of 2 edges of the WordNet graph. In particular we consider only the graph made by hypernym and hyponym relations for the nouns and only by the hypernym relations for the verbs. Then, in order to compute the distance between two input words we just use the Jaccard Index on the obtained sets.
- *Strategy 3 (S3)* based on the *average polarity similarity*, that takes into account if two adjectives belong to the same polarity region, i.e positive, negative or objective. To compute this measure, we use the SentiWordNet resource [3]. In more detail using the knowledge of our resource, we divide the polarity region in tree equal subspaces: positive, negative and objective. Given an adjective, we retrieve all its synsets from the SentiWordNet resource. Then, we classify each retrieved synset with a *local polarity indicator* based on the thresholds used in the subspaces definition. Then, we define a *global polarity indicator* for this adjective as the most common local polarity indicator among all of its synsets and we also associate to it a *global polarity measure* computed as the average values among the ones that belong to the same space of the global polarity indicator. Now, we compute the average polarity similarity between two adjectives as the minimum global polarity measure between the two words if both the adjective has the same global polarity indicator otherwise it is 0.

In Table 2, we depicted the strategy used to compute the distances among words according to their PoS label. We use the enumeration introduced in this section to represent the selected approach, 0 means that we choose to not compute any distances. We note also that in the case of adjective that are collapsed with their negation we apply the strategy 3 if also the other adjective was in the same situation. Now, for example let us consider two gram elements $g_1$,$g_2$ $\in \mathcal{E}^3$. The similarity $sim(g_1, g_2)$ is computed following the strategies defined in Table 2 for each couple of words obtained from words in $g_1$ and in $g_2$, then the average value is returned. In particular, we note that just for gram elements belonging to $\mathcal{E}^2$ and $\mathcal{E}^3$ we consider all the possible couples. We note that also this approach can take advantage of some caching operation on all the resources involved.

[2]We confider each belief against the rest as in a leave-one-out approach.

| | Noun | P. Noun | Adjective | Verb | Adverb |
|---|---|---|---|---|---|
| Noun | S2 | 0 | 0 | 0 | 0 |
| P. Noun | 0 | S1 | 0 | 0 | 0 |
| Adjective | 0 | 0 | S3 | 0 | 0 |
| Verb | 0 | 0 | 0 | S2 | 0 |
| Adverb | 0 | 0 | 0 | 0 | S1 |

**Table 2: Strategies used to compute the distances among words based on PoS tagging. The enumeration is the one described in Section 6. P.Noun means Proper Noun**

# 7. MINING THE COGNITIVE FEATURES

In this section we describe how we process a set of resources in order to detect the introduced Cultural Models. Let us suppose that we have a network of resources such as web pages. Each resource can be automatically processed in order to extract useful information such as images, tables, text content, links, html structures. Let us explain how we process the text data. We design a process flow, which is depicted in Figure 4, that is based on four main modules: *Resource Knowledge Processing*, *Cultural Knowledge Processing*, *Context Selection* and *Cultural Model Detection*. The Resource Knowledge Processing module has the aim to extract all the Text Signals from an input resource, as described in Section 4. For example, it takes as input resource a document $d$ and it returns a structure made by a sentence $s_i$ belonging to $d$ and some useful Text Signals extracted from $s_i$ such as $\langle s_i, \mathcal{E}_{s_i}, \mathcal{E}^2_{s_i}, \mathcal{E}^3_{s_i}, \mathcal{S}^2_{s_i}, \mathcal{S}^3_{s_i}\rangle$. The Cultural Knowledge Processing module has the aim to derive the Cognitive Annotations related to a selected Cultural Model. It takes in input a Cultural Model $\mathcal{M}=\{e_i,\ldots,e_n\}$ and for each $e_i \in \mathcal{M}$ it retrieves a Cognitive Annotation. It chooses for each $e_i$ the most important Cognitive Annotation using the reliable measures computed in the processing described above. It returns for each $e_i \in \mathcal{M}$ a structure such as $\langle e_i, \mathcal{E}_{e_i}, \mathcal{E}^2_{e_i}, \mathcal{E}^3_{e_i}, \mathcal{S}^2_{s_i}, \mathcal{S}^3_{s_i}\rangle$ where the sets are the union of the sets of gram elements of the same gram element type belonging to the same Cognitive Annotation. For example $\mathcal{E}^3_{e_i}$ is the union of all the tri-grams that belong to the Entity Text Cognitive Annotation selected to be representative for $e_i$. The *Context Selection Module* has the aim to choose some group of sentences that are indicative of our further analysis. The Cultural Model Detection, instead, has the aim to evaluate the presence of each Cognitive Feature in the initial resource. In this way we are able to understand if the considered Text Signals have the cognitive signatures related to the selected Cultural Model. Let us give more details about the last two modules in the following subsections.

## 7.1 Context Selection Module

In this module, we start to consider a different granularity for our analysis, in particular we define the *Context* as set of subsequent sentences. At this stage the initial resource, for example a document, can be seen as $\mathcal{C}=\{c_1,\ldots,c_m\}$, being its generic element $c_i=\{\langle s_{i-k}, \mathcal{E}_{s_{i-k}}, \mathcal{E}^2_{s_{i-k}}, \mathcal{E}^3_{s_{i-k}}, \mathcal{S}^2_{s_{i-k}}, \mathcal{S}^3_{s_{i-k}}\rangle, \ldots, \langle s_{i+k}, \mathcal{E}_{s_{i+k}}, \mathcal{E}^2_{s_{i+k}}, \mathcal{E}^3_{s_{i+k}}, \mathcal{S}^2_{s_{i+k}}, \mathcal{S}^3_{s_{i+k}}\rangle\}$ a Context of $2k+1$ subsequent sentences, with $k<i$ together with all the text signals extracted for each sentence . In this module we define a filter able to select only the Context that we need to process by a next module. The filter is designed as a statistical decision process based on the analysis of the uni-gram of each Context. In particular we model the relevance of the Context in terms of trails of a binary random

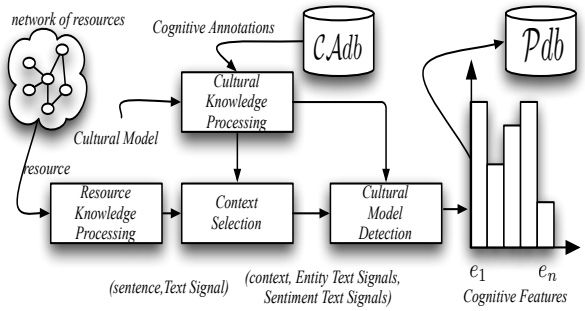


**Figure 4: The modules used in our process.**

variable (r.v). In fact, we map each uni-gram belonging to a context $c_i$ to an independent and identically distributed (i.i.d) binary r.v and we define a kind of *relevance* of the input Context through a Bernoulli process. Let us consider the set $\mathcal{E}_{c_i}$ made by the union of the different $\mathcal{E}_{s_*}$ being $s_*$ a sentence belonging to $c_i$ and $N_{c_i}$ its cardinality. We define a grade of relevance $r$ as $r$ success in $N_{c_i}$ trial as follows:

$$P(r|N_{c_i},\theta) = \binom{N_{c_i}}{r}\theta^r(1-\theta)^{(N_{c_i}-r)} \quad (3)$$

For our decision problem, we are interested in the Bayesian estimation of $\theta$. This operation is done using as "observed trials" the $N_{c_i}$ words. Let us now explain how we map the gram elements in a set of binary variables. We transform each uni-gram $g \in \mathcal{E}_{c_i}$ in a sequence of relevance ($ro$) or not relevance ($nro$) observation through the function $f_r$ defined as follows:

$$f_r(g,\mathcal{M}) = \begin{cases} ro \text{ if } g \in \mathcal{E}_{\mathcal{M}} \\ ro \text{ if } \exists\ g^* \in \mathcal{E}_{\mathcal{M}} : dist(g,g^*) < \epsilon \\ nro \text{ otherwise.} \end{cases} \quad (4)$$

Being $\mathcal{E}_{\mathcal{M}}$ the union of all the uni-gram attached to the $e_i \in \mathcal{M}$, with $\mathcal{M}$ the chosen Cultural Model, *dist* a distance between two uni-grams computed according to the strategy defined in Section 6 and $\epsilon$ a fixed threshold. We use for the estimation of $\theta$ as prior an non-informative beta distribution $Beta(\theta|\alpha,\beta)$, with $\alpha=\beta=0.5$ . According to the Bayesian Analysis, the estimator $\hat{\theta}$ has a distribution $Beta(\alpha+n^{ro},\beta+n^{nro})$, being $n^{ro}$ and $n^{nro}$ the number of times we observe a relevance or a not-relevance sample respectively. Now the selection process is computed as follows: i) we first select those contexts for which the base condition $E[\hat{\theta}] \geq 0.5$ is verified; ii) then we send to the next module those contexts whose discriminative measure of relevance ($dr$) exceeds a given thresholds. The $dr$ is defined as follows:

$$dr = \frac{E[\hat{\theta}]-0.5}{\sigma(\hat{\theta})} \quad (5)$$

being $\sigma(\hat{\theta})$ the standard deviation of $\hat{\theta}$.

## 7.2 Cultural Model Detection Module

In this module, we propose a model able to evaluate the diffusion of the Cognitive Features on the input resources by analysing the extracted Text Signals in order to extract cultural evidence from them. First, we traduce all the selected

Contexts in the previous module as binary strings using the Cognitive Annotations extracted from the selected Cultural Model. This process is made by a function $f$ defined as follows:

$$f(\mathcal{A}^*_{e_i}, \mathcal{A}^*_{c_j}) = \begin{cases} 1 \text{ if } \mathcal{A}^*_{e_i} \cap \mathcal{A}^*_{c_j} \neq \emptyset \\ 1 \text{ if } \exists g^*_{c_j} \in \mathcal{A}^*_{c_j}, g^* \in \mathcal{A}^*_{e_i} : dist(g^*_{e_i}, g^*_{c_j}) < \epsilon^*_{e_i}, \\ 0 \text{ otherwise.} \end{cases} \tag{6}$$

Being $\mathcal{A}^*_{c_j}$ a set of text signal extracted from the sentences belonging to a $c_j$ that was selected in the previous module and $\mathcal{A}^*_{e_i}$ the set of gram belonging to the selected Cognitive Annotation. For example $\mathcal{A}^*_{c_j}$ can be $\mathcal{A}^{\mathcal{S}^3}_{c_j} = \mathcal{S}^3_{s_1} \cup \ldots \cup \mathcal{S}^3_{s_{n_{c_j}}}$, where $\{s_1, \ldots, s_{n_{c_j}}\}$ are the sentences belonging to $c_j$. We note that in this phase we only considered the bi-grams and tri-grams. Then, we have *dist* and $\epsilon^*_{e_i}$ that are distances computed as describe in Section 6 and a fixed threshold related to the text element type and Cognitive Feature, respectively. We note that this distance requires that the input grams are of the same gram element type. This means, for example, that we compare a tri-gram extracted from an Entity Text Cognitive Annotaion with a tri-gram coming from the Text Entity Signal of a Context. For sake of clarity, we depicted in Figure 5 how we generate these binary signals from a generic Context $c_i$. To evaluate how the Cognitive Features are spread around the Contexts, we considered a hierarchical bayesian model that can be seen as the generative models of these binary strings.

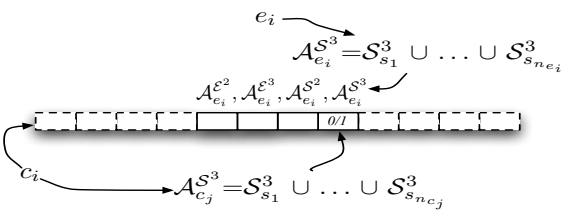


**Figure 5: How we build the binary strings from the Context $c_i$ using the approach described in Section 7.2.**

These generative models are well studied in statistical natural language processing to inference topic distributions on corpus. Our approach has some similarity with the graphical model proposed in [6]. The model is depicted in Figure 6 using the Plate notation [6]. In particular in this generative model, as described in Figure 6, we have a document $d$ described by a set of Context $\mathcal{C}$. From $d$ we sample with uniform distribution a Context $x$. Then for each Context $x$, we sample a Cognitive Feature $z$ from its set $\mathcal{CE}$ with a multinomial distribution with parameters $\Theta$. Then, from each Cognitive Feature $z$ we sample a binary variable $w$ from a binomial distribution with parameter $\phi$ four times. In this setting we have a learning problem, where we use the Bayesian theory. This means that we want to estimate the distribution of latent variables using some data and prior over these latent variables. In particular we use as prior the Dirichlet distribution $\alpha$ for the multinomial distribution and the Beta distribution ($\beta$) for the Binomial distribution. As data to observe we use the binary strings obtained from

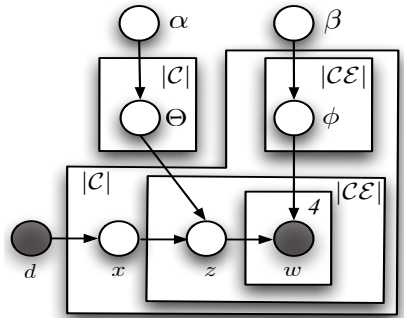


**Figure 6: Hierarchical Bayesian Model used to estimate the distribution of Cognitive Features in a document d. The latent variables are represented by white nodes.**

the different Contexts coming from a document $d$, that were selected by the previous module. We are interested to estimate the distribution $\Theta$ that gives the information of how the Cultural Features are spread around the Context. To estimate this distribution we use some equations used in the well known Collapsed Gibbs inference algorithms [6], that is typically used to estimate the latent variables in Bayesian graphical model. We note that we use this approach to measure a distribution of r.vs rather than to classify new data by training a bayesian learner.

# 8. CASE STUDY

We apply our framework in the domain of the Islam religion, with the aim to understand how the beliefs introduced in Table 1 are spread around three main population: Moderate Arab, Moderate USA, Extreme. The data was collected from web sites in Arabic and English language, in particular they are collected using Google search engine both in English and in Arabic with some keywords related to our beliefs. In particular, we collected 80 documents, 23 of them were used to select some Cognitive Annotation for each belief defined in Table 1 and the others to run our experiments. The domain experts report for each belief a set of sentences that are used as Cognitive Annotations. We note that the Cultural Model considered in our case study is made only of Cognitive Features such as $\langle B, -, \delta \rangle$ or $\langle B, v, \delta \rangle$. In other words, we consider as a Cognitive Feature a belief with its values ($\langle B, v, \delta \rangle$) for the Context Selection module and the belief without its values ($\langle B, -, \delta \rangle$) for the Cultural Model Detection. We do not take into account an initial measure of the attitude of each belief so we start our process with $\delta = 0$. We compute also for each element in the Cognitive Annotation a reliable measure, which is a value in the interval of $[0, 1] \cap \mathbb{R}$ as described in Section 5. We select as Context a fixed group of 5 sentences, and we choose the following thresholds $\epsilon$=0.8 and $dr$>0.8 for the Context Selection module. For the Cultural Model Detection module, we use 0.8 for each $\epsilon^*_{e_i}$. We note also that for the document written in the Arabic language, we first run some machine translation procedure and then these documents were corrected by a native arabic speaker in order to overcome the problem related to the imperfection of the machine translation algorithms. In Table 3 is depicted the data about the dimensions of our collection and the average numbers of

| | Num. of Documents | AVG Num. of Contexts | AVG Num. of selected Contexts |
|---|---|---|---|
| Moderate USA | 24 | 856 | 43 |
| Moderate Arab | 18 | 623 | 66 |
| Extreme | 15 | 739 | 36 |

**Table 3: Summary of the collected data**

| | C-KM | R-KM | C-KC | R-KC |
|---|---|---|---|---|
| Moderate USA | 22 | 0.4 | 32 | 0.4 |
| Moderate Arab | 16 | 0.7 | 40 | 0.6 |
| Extreme | 22 | 0.8 | 12 | 0.6 |
| | C-CH | R-CH | C-CD | R-CD |
| Moderate USA | 9 | 0.7 | 34 | 0.7 |
| Moderate Arab | 14 | 0.7 | 48 | 0.8 |
| Extreme | 31 | 0.8 | 12 | 0.7 |
| | C-IES | R-IES | C-IEI | R-IEI |
| Moderate USA | 16 | 0.7 | 23 | 0.5 |
| Moderate Arab | 13 | 0.7 | 27 | 0.4 |
| Extreme | 22 | 0.6 | 17 | 0.4 |
| | C-JA | R-JA | C-JI | R-JI |
| Moderate USA | 19 | 0.5 | 33 | 0.7 |
| Moderate Arab | 22 | 0.8 | 42 | 0.6 |
| Extreme | 25 | 0.8 | 15 | 0.6 |

**Table 4: Summary of the results in our case study. The full names of * in C/R-{*} are depicted in Table 1.**

selected contexts after running the Context Selection process. The results of this case study are described in Tables 4 and in Figure 7. In Table 4, there are represented the number of contexts (C-{*}) and the average value of the reliable measure (R-{*}) that is assigned to each Cognitive Feature. Each row in Table 4 is related to a member of the considered population. Eventually, in Figure 7, we have the average values of context assigned to each belief in log scale. As time performance, we note that for an average number of 700 contexts, 8 Cognitive Features, and 2000 iteration of Gibbs sampler, we obtained an estimation of Cognitive Features distribution in about 3 hours of wall-clock time on standard 3GHz 4GB RAM PC workstation. We note that Cognitive Features such as Coherence Homogeny and Diversity are good indicators of the cultural differences among our population, this is also justified by the better annotation that the domain experts did for those beliefs as suggested by the average values of the reliable measures associated to them.

# 9. CONCLUSION

In this paper we have presented a general framework to analyse cultural behaviour on text data. In particular we propose a methodology based on concepts such as Cognitive Features, Text Signals and Cognitive Annotations. We also proposed some computational methods in order to use our framework with some text data. These computational methods come from the area of Bayesian Learning. In particular we designed a graph model used to estimate the diffusion of cultural beliefs within a population. A Case Study in the extreme religious domain was also reported with some results. In this Case Study, we can see how the Cognitive Features can be used to discriminate among different population from a cultural perspective.

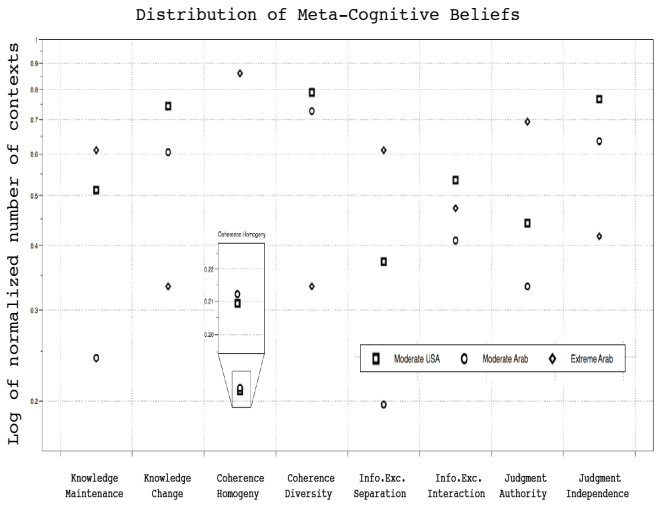


**Figure 7: How the Cognitive Features (CFs) are spread over the different populations in this case study. In particular, we have on x-axis the CFs and on y-axis the log of normalized number of Contexts.**

## 9.1 Acknowledgments

This research was supported by Contract N00014-10-C-0078 from the Office of Naval Research.

# 10. REFERENCES

[1] M. W. Berry and M. Castellanos. *Survey of Text Mining II: Clustering, Classification, and Retrieval*. 1 edition, 2007.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal Machine Learning*, March 2003.

[3] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *International Conference on Language Resources and Evaluation*, 2006.

[4] C. Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, 1998.

[5] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[6] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. *ACM Transaction Information System.*, January 2010.

[7] W. Sieck. Metacognition and religion: The case of islamic extremism. In *Annual Meeting of the International Association for the Cognitive Science of Religion, Boston, MA*, 2011.

[8] W. Sieck, L. Rasmussen, and P. R. Smart. Cultural network analysis: A cognitive approach to cultural modeling. *Network Science for Military Coalition Operations*, 2010.

[9] J. Yang, M. Morris, J. Teevan, L. Adamic, and M. Ackerman. Culture matters: A survey study of social qa behavior. In *International AAAI Conference on Weblogs and Social Media*, 2011.