# DeCAT 2015 - Workshop on Deep Content Analytics Techniques for Personalized and Intelligent Services

Lora Aroyo[1], Geert-Jan Houben[2], Pasquale Lops[3], Cataldo Musto[3], and Giovanni Semeraro[3]

[1] Department of Computer Science, VU University Amsterdam, The Netherlands
[2] Delft University of Technology (TU Delft), The Netherlands
[3] Department of Computer Science, University of Bari "A. Moro", Italy

## 1 Introduction

According to a recent claim by IBM, 90% of the data available today have been created in the last two years. This uncontrolled and exponential growth of the online information gave new life to the research in the area of user modelling and personalization, since information about users preferences, sentiment and opinions can now be obtained by mining data gathered from many heterogeneous sources.

As an example, many recent work rely on the analysis of the content posted by people on social networks and micro-blogs to unveil latent information about their interests, automatically extract people personality traits, build preferences models on the ground of textual reviews, and so on. At the same time, the recent phenomenon of (Linked) Open Data fueled this research line by making available a huge amount of machine-readable textual data.

All these trends paved the way to the design of intelligent and personalized systems able to extract some real value from this plethora of rough textual content produced on the Web: examples of such services are online brand monitoring platforms, social recommender systems and smart cities-related applications, as incident detection systems or personalized city tour planners.

However, a complete exploitation of such textual streams requires a comprehension of the information conveyed by people. In turn, this requires a deep understanding of the language, which is not trivial. The major goal of this workshop is to stimulate the attention of the scientific community on the aforementioned topics. The workshop aims to provide a forum for discussing open problems, challenges and innovative research approaches in the area, in order to investigate whether the adoption of techniques for semantic content representation and deep content analytics can be effective to build a new generation of intelligent and personalized services based on the analysis of Social, Big and Linked Open Data.

## 2 Motivations and Workshop Topics

The importance of user modeling and personalization is taken for granted in several scenarios. According to this widespread paradigm, each user can be modeled to some (explicitly or implicitly gathered) information about her knowledge or about her preferences, in order to adapt the behavior of a generic intelligent system to her specific characteristics.

However, the rapid growth of social networks changed the rules for personalization, since the spread of these platforms radically changed and renewed many consolidated behavioral paradigms. Indeed, people today exploit these platforms for decision-making related tasks, to support causes, to provide their circles with recommendations or even to express opinions and discuss about the city or the place where they live. Thanks to the heterogeneous nature of the discussions that take place on social networks, a lot of new data are continuously available and can be gathered and exploited to build richer and more complete user models, to discover latent communities, to infer information about users emotions and personality traits, and also to study very complex phenomena, such as those related to the psycho-social sphere, in a totally new way. At the same time, thanks to crowdsourcing, a huge amount of content-based information has been made available in open knowledge sources as Wikipedia and the Linked Open Data Cloud.

Given that most of the information stored in these modern data sylos is made available as textual content, a consequence, a complete exploitation of these rich information sources requires a big effort on the definition of models and techniques able to effectively process the content and to represent it in a machine-readable form, in order to unveil the latent semantics and trigger more effective personalization and adaptation pipelines. This is not a trivial task, since this process requires a deep comprehension of the language, which in turn typically requires a combination of techniques coming from Machine Learning and Natural Language Processing areas.

The main goal of the workshop is to stimulate the discussion around problems, challenges and research directions regarding the exploitation of content-based information sources (Big, Social and Linked Data) for personalization and adaptation task and to foster the design of a new generation of intelligent user-centered services.

We hope the workshop will stimulate discussions around the presented papers, the invited talk and the following questions:

- What is the impact of semantics in personalization and adaptation tasks?
- Can social media improve the representation of user interests?
- Can semantic analysis technique improve the representation of user interests?
- Can these data sylos (Wikipedia, DBpedia, Freebase) be useful for personalization and adaptation tasks?
- Which data sylos are more effective to model user interests and preferences?
- What content-based information is more useful to personalize and adapt the behavior of modern intelligent systems?

- Does a semantic representation of the information improve the effectiveness of personalization tasks?
- Does a semantic representation of the information improve the transparency of such platforms?
- Can the analysis of content coming from social media provide some information about user personality traits?
- How do people deal with privacy issues? Are them willing to trade better personalization with a larger tracking of their activities on the Web?
- Is it possible to think about a novel generation of adaptive platforms able to completely exploit all the available information?

## 3 Contributions

Five papers will be presented in DeCAT 2015. The papers were accepted after a peer-review process: each paper was reviewed by at least two members of the Program Committee and evaluated in terms of Significance, Technical Quality and Novelty of the approach.

In their contributions, Abela et al. [1] tackle the Personal Information Management (PIM) problem, and propose a methodology to automatically organise personal information accessed by the user into task-clusters. To this aim, the authors transparently exploiting the users behaviour while performing some tasks. A distinguishing aspect of their work is the usage of PiMx app. a tool which can be of interest for other researchers working on task clustering.

Next, Papadopoulos et al. [2] present ongoing work on the formalization of a persons creativity, modelling it in terms of four characteristics of the personal content creations, namely novelty, surprise, rarity and recreational effort. Based on such formalization, the paper also presents the Creativity Profiling Server (CPS), a system implementing the aforementioned user modelling framework for computing and maintaining creativity profiles

The analysis of social media is the focus of the work proposed by Matta et al. [3]. In this paper the authors perform an interesting analysis of the connection between Bitcoin's price and the volume of Tweets about the topic. Specifically, the authors use an external API to crawl Twitter data and assign a sentiment to it. Next, they analyze how the price of Bitcoins changed over time and they looked for some connections between these aspects. A thorough analysis of the time series showed that some connection (calculated as the cross-correlation between time series) exists.

In the only short paper accepted, Pentel investigated the relation between reading and writing skills in the task of age-based categorization. In this contribution [4] he presents results of a study on age-based categorization of short texts as 85 words per author. He introduced a novel set of features that will reliably work with short texts, which makes easy to extract from the text itself without any outside databases.

Finally, Basile et al. [5] propose a content-based and time-aware movie recommendation approach. The novel contribution is the time-adaptivity for a content-based technique. The authors proposed an approach that models short-term

preferences by adopting a content-based sliding window approach: when a new ratings comes into the system, the replacement of an older one is performed by taking into account both a decay function for user interests and content similarity between items on which ratings are provided, computed by distributional semantics models. The authors carried out an evaluation that demonstrate that their approach overtake the baseline FIFO strategy.

## References


1. Abela, C., Staff, C., Handschuh, S. *Automatic Task-Cluster Generation based on Document Switching and Revisitation.* In Proceedings of DeCAT 2015 - 1st Workshop on Deep Content Analytics Techniques for Personalized and Intelligent Services, co-located with UMAP 2015, Dublin (2015).
2. Papadopoulos, G., Karampiperis, P., Koukourikos, A., Konstantinidis, S. *Creativity Profiling Server: Modelling the Principal Components of Human Creativity over Texts.* In Proceedings of DeCAT 2015 - 1st Workshop on Deep Content Analytics Techniques for Personalized and Intelligent Services, co-located with UMAP 2015, Dublin (2015).
3. Matta, M., Lunesu, M.I., Marchesi, M. *Bitcoin Spread Prediction Using Social And Web Search Media.* In Proceedings of DeCAT 2015 - 1st Workshop on Deep Content Analytics Techniques for Personalized and Intelligent Services, co-located with UMAP 2015, Dublin (2015).
4. Pentel, A. *Employing Relation Between Reading and Writing Skills on Age Based Categorization of Short Estonian Texts.* In Proceedings of DeCAT 2015 - 1st Workshop on Deep Content Analytics Techniques for Personalized and Intelligent Services, co-located with UMAP 2015, Dublin (2015).
5. Basile, P., Caputo, A., de Gemmis, M., Lops, P., Semeraro, G. *Modeling Short-Term Preferences in Time-Aware Recommender Systems.* In Proceedings of DeCAT 2015 - 1st Workshop on Deep Content Analytics Techniques for Personalized and Intelligent Services, co-located with UMAP 2015, Dublin (2015).

# Automatic Task-Cluster Generation based on Document Switching and Revisitation

Charlie Abela[1], Chris Staff[1], and Siegfried Handschuh[2]

[1] Department of Intelligent Computer Systems,
University of Malta, Malta
{charlie.abela,chris.staff}@um.edu.mt
[2] Department of Computer Science and Mathematics,
University of Passau, Bavaria, Germany
{siegfried.handschuh}@deri.org

**Abstract.** Personal Information Management (PIM) research is challenging primarily due to the inherent nature of PIM. Studies have shown that people often adopt their own schemes when organising their personal collections, possibly because PIM tool-support is still lacking. In this paper we investigate the problem of automatic organisation of personal information into task-clusters by transparently exploiting the user's behaviour while performing some tasks. We conduct a controlled experiment, with 22 participants, using three different task-execution strategies to gather clean data for our evaluation. We use our PiMx (PIM analytix) framework to analyse this data and understand better the issues associated with this problem. Based on this analysis, we then present the incremental density-based clustering algorithm, iDeTaCt, that is able to transparently generate task-clusters by exploiting document switching and revisitation. We evaluate the algorithm's performance using the collected datasets. The results obtained are very encouraging and merit further investigation.

**Key words:** Density-Based Clustering, Personal Information Management, Task Clusters

## 1 Introduction

When we are performing some task on our desktop, we tend to spend a considerable amount of time looking back, establishing past references and remembering [10]. Whether performing the task requires us to search for information on the Web, reply to some email we've received, or resume writing some other document which we've worked on the day before, we tend to rely on our organisational skills and the support of search, bookmarking and history tools [9].

The common feature in these tools is their ability to help us find or re-find information by exploiting revisitation [10]. However, most of these tools tend to consider the user's information-seeking activities as unrelated events, unlike the way we actually organise things, which is usually in terms of directories (on our

desktop) and tasks (conceptually) [11, 13]. Furthermore, humans are by nature or by force multi-taskers, and tool-support should cater for situations whereby users switch between one task and another or are interrupted [4, 11].

In this paper we align our research with efforts such as those of [12, 1] and investigate how to transparently cluster documents such as Web-browsed documents, office-related documents and emails, which are viewed by the user and belong to the same task.

As documents are presented to users in windows and tabs, and users switch between them, we collect and exploit evidence of the users *"window switching behaviour"* to identify and generate task-clusters[3]. These clusters can be used to re-find specific task-related documents and to resume a task, if or when, the user is interrupted. Currently, we do not label the identified task-clusters and we are not considering the content of the accessed documents as was the case in [12].

We adopt an unsupervised method that treats the accesses to documents as an undirected *activity graph*, $G_a(V, E)$, based on which we create the task-cluster's graph $G_c(V, E)$, which is both weighted and undirected. The edge weight $w(u, v)$ in $G_c(V, E)$ reflects the strength of the association between two nodes $u$ and $v$. In the text we tend to use interchangeably the words "document" and "node", depending on the perspective we consider.

We performed a controlled experiment, conducted with 22 participants, using 3 different task-execution strategies to gather user's window-switching behaviour data for our evaluations. We separate the collected data into 3 groups depending on the task-execution strategy adopted: in succession with no interleaving (used as baseline), interleaved and interleaved over different sessions. We also developed the PIM analytix *PiMx* framework through which it was possible to simulate, off-line, the task-execution over the collected data and at the same time exploit network-analytics to analyse and visualise $G_a(V, E)$ and $G_c(V, E)$. Through PiMx we were able to understand better which algorithmic approach was more suitable.

In our approach we have factored-in an important feature, which to our knowledge has not been addressed yet. We refer to the incremental nature of the task and task-clusters, with nodes and edges being added over time as the user visits and re-visits documents. We extend the work of [6], and propose an incremental density-based clustering algorithm, which we call **iDeTaCt** that identifies the dense regions from the less denser ones in the accessed documents' space, identifying the task-clusters in the process.

We evaluated our approach to verify how well our algorithm is able to: (i) identify those nodes that belong together in a task-cluster and (ii) identify when a switch between two nodes is effectively a task-switch. The results are very encouraging and **iDeTaCt** managed to cleanly separate all the tasks, using the data of all the participants from the baseline group which performed the tasks in succession. When we used the data with interleaved tasks, **iDeTaCt** was

[3] A task-cluster is a group of documents that pertain to a task.

found to be sufficiently reliable in more than 50% of the cases, at the expense of capturing less documents from the referenced tasks.

The rest of the paper is structured as follows. In Sec. 2 we present research which is closely related to our own. We follow up with a detailed description of the controlled experiment that we conducted to collect reliable data. In Sec. 4 we give an overview of the PiMx framework which we use to analyse the collected data. We introduce our *iDeTaCt* algorithm in Sec. 5, which is followed by the evaluation and future work sections.

## 2 Related Work

In our work we draw parallels with research related to task-switching and identification, and others that have adopted the graph data structure as the underlying representation for the user's switching and revisitation behaviour.

In his thesis [10], Mayer presented an integrative history, visualization tool entitled SessionGraphs. The tool allows a user to view her browsing activity as an animated, interactive graph and to organise the visualisations according to her tasks. We have adopted a similar graphical approach for our PiMx framework, however the scope behind PiMx was that of allowing a researcher to better understand the issues related to the problem of automatic task-cluster generation rather than to provide browsing support.

The search bar presented in [11] persistently maintains a hierarchical Web history organised around search topics and queries. It assists users in organising complex searches and re-acquiring the context of a suspended search, based mainly on topic and query-driven groupings. The multitasking bar presented by [13] copes with both multiple tasks as well as multiple session tasks. They considered a task as having different states and it was up to the user to maintain and organise a task. Our approach aims to transparently automate the task-cluster generation without requiring any user intervention.

A semi-automated approach to task-identification was adopted by [3] which used activity-log analysis to group the accessed resources based on cues generated by an individual while performing some task. We adopt this same approach to collect the user's activity information through dedicated application plug-ins. We maintain a global desktop history with information about all the created, accessed and edited documents which also includes Web-related documents and queries.

To identify documents pertaining to a task [12] computed document content similarity and applied a maximal clique-finding algorithm over the user's switching activities. Unlike this approach, we currently do not intend to exploit the content of the accessed documents, however we will consider the incremental characteristics underlying an information-finding process.

In [1] a PageRank-like association heuristic was used to compute the association between windows opened on the desktop and presented a visualisation through which windows that were frequently clicked in sequence, were displayed closer together. However, no task-clusters were explicitly generated.

In [7, 8] an incremental density-based graph clustering approach is used to cluster documents and find interesting subgraphs, respectively. Density-based clustering is quite interesting since it is capable of coping effectively with noise. In our case this will be a major challenge since users tend to constantly switch between tasks, with the result that it would be more difficult to deal with those documents that are accessed in between tasks.

## 3 Controlled Data Collection Experiment

Evaluating PIM related research is inherently difficult, in particular due to the lack of readily available datasets. We therefore conducted a data-collection experiment in a controlled environment, to collect clean data related to the window-switching behaviour of the participants while performing some predefined tasks.

We set up a cluster of machines in one of our laboratories, each running Windows OS and having two activity-monitoring applications installed on them. One of the applications monitored browsing activity on Firefox[4] while the other monitored file browsing activity (e.g. of word processing documents) on the desktop. The participants were advised about this monitoring and were assured that the data would be anonymised and used only for the specified research purpose before the start of the experiment. This consisted of all the participants performing the same three, predefined information-seeking tasks, by answering a number of questions related to specific topics. Apart from seeking out information, participants had to compile a document with the relevant answers for each task, which they had to email to us at specified intervals. Our methodology was in line with that adopted by [11, 10].

The tasks required participants to provide specific information about the planning of a *vacation* in a specific country; answering questions related to the research area of *human computation*; and providing information about any two upcoming *music events*. The tasks were conducted either over single or multiple sessions. At pre-established intervals, unknown to the participants, we sent out emails that either informing them what a task entails or else requesting that they switch to another task.

There were in total 22 participants, 25% of whom were female. The participants were students and members of staff (lecturing and administration) from the Faculty of ICT within the University of Malta. The students were compensated €10 for their participation in the experiment.

The participants were split into three groups. Each group performed the experiment separately from the others. The groups were split as follows:

i. *Group 1*: the 7 participants in this group completed each of the three tasks in sequence, starting with task 1, followed by tasks 2 and 3, without interruptions. We use the data from this group as the baseline for our algorithm, since the tasks are clearly separated from each other;

[4] `https://www.mozilla.org/en-US/firefox/desktop/`

ii. *Group 2*: there were 10 participants in this group. They performed the three tasks in a single session. They started working on task 1 but were interrupted with an email from us requesting that they start task 2. After some more time we sent another email requesting the participants to stop working on task 2 and start working on task 3. We later interrupted them with yet another email requesting that they switch back to task 2, finish it, and then switch to, and complete tasks 1 and 3, in this order. In this way the tasks were interleaved and thus identifying which documents pertained to which task, becomes even more challenging;
iii. *Group 3*: the 5 participants in this group performed the tasks in the same order as Group 2 and with similar interruptions, however they were stopped 30 minutes into the session. Later on we asked them to continue the experiment in another session, which took place some days later. During the second session they had to resume the tasks and complete them in a sequential order with no further interruptions. In this way we wanted to introduce some more challenges to the participants, since they had to remember what they had been working on and recall the state of the task/s before they were stopped.

Although we tried to have an equal number of participants in each group, due to availability issues of our participants we had to somewhat relax this aspect. Furthermore, the data collected from one participant from Group 1 and another from Group 2 was unusable due to issues with the data-logging applications, which were unfortunately, not noticed in time.

The logged data included information about the type of event (e.g. navigational and tabbed events), the application that generated the event, the timestamp, the URL of the document accessed as a result of the event, an excerpt of text from the window caption. Other information, specific to particular events was also captured. This included, the URL of the page that was in focus before the event was triggered, as is the case of the navigational events. We also captured the file name and whether a document was edited or not, in the case of the desktop's file-related events. The data logged from each participant was anonymised and cleaned for further processing.

## 4 PiMx: tool for analysing the data

We implemented a tool, called PiMx (**P**ersonal **i**nformation **M**anagement analyti**x**) to analyse the collected data and understand better how we can algorithmically exploit document switching and revisitation to generate the task-clusters.

*PiMx* allows a researcher to load a user's activity-log and to simulate the execution of the task-trail for that user. This process can be paused and resumed at any time, allowing the researcher to analyse and compare the evolving task-trail through different views, see Fig. 1. The *PiMx-History* is similar to the tool developed by [5] and allowed us to view details related to all accessed documents, including the URI and amount of revisitations. It is also possible to filter the

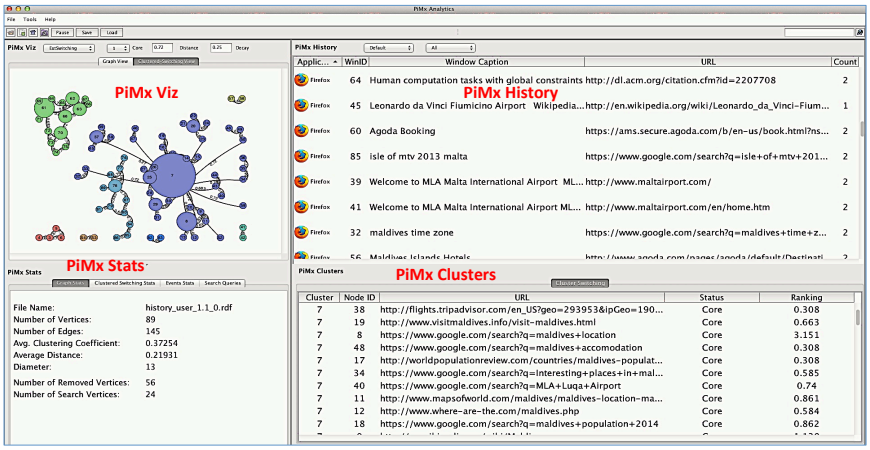


Fig. 1: PiMx Interface

data by different time windows (e.g. last hour, last 4 hours, today, yesterday etc.), as well as by application and file-type.

The *PiMx-Viz* is an interactive component inspired by the SessionGraph described in [10]. This provides visualisations of the unadulterated activity-graph as it evolves over time and the task-clusters as they are generated by the applied algorithm. The size of the nodes is relative to the number of accesses and it is also possible to click on each node separately and to visualise the induced subgraph generated by the nodes' neighbourhood.

The *PiMx-Stats* component was inspired by graphical tools such as Visone[5] and Gephi[6]. It presents a number of graph related statistics, such as the number of vertices and edges, the clustering coefficient, the average distance and diameter of the graph. There is also information about the number of search and removed nodes. Search nodes represent the pages associated with a search engine query. The relevant query and the number of times that this search node was accessed are displayed. Information about the type and number of occurrences of the events that were triggered is also provided.

Through the *PiMx-Clustering* view it is possible to view the details of the documents pertaining to each task-cluster. Each cluster is assigned a unique ID and each document in a cluster has associated with it a ranking value and information about the status generated by the algorithm. More details about this algorithm are found in Sect. 5

## 5 Incremental Graph Clustering Approach

In this section we give an overview of the incremental density-based task clustering approach that we've adopted. The scope behind our algorithm **iDeTaCt**

[5] `http://visone.info/`
[6] `http://gephi.github.io/`

is two fold: (i) identify those nodes that belong together in a task-cluster and (ii) identify when a switch between two nodes is effectively a task-switch.

Clustering entities into dense parts allows for the discovery of interesting groups in different networks. Furthermore, clustering on time-evolving networks is still an open research problem that has been addressed through different approaches including incremental clustering [2] which tracks the granular dynamics of a network, such as edge and node addition and deletion, rather than a time-window. This approach is quite applicable to the dynamics of information-seeking behaviours, whereby new documents are added over time, which in turn need to be assigned to an existing or new task-cluster.

### 5.1 iDeTaCt: incremental Density-based Task Clustering

The density-based clustering algorithm DBSCAN proposed by [6] produces partitional clustering, whereby a cluster is considered to be a continuous area of arbitrary shape that is denser than its surroundings. DBSCAN relies on the idea that the neighbourhood of a node up till some given radius $\epsilon$ defines the *"importance"* of that node. Nodes that have a minimum number, $\eta$, of other nodes at a distance less then $\epsilon$ are termed as *core nodes*. On the other hand, a node that has no such neighbourhood is given the status of *noise node*, unless it is contained within the neighbourhood of a core node, in which case it is assigned the status of a *border node*. Thus $\epsilon$ and $\eta$ ensure that node neighbourhoods are dense areas.

In our case, we consider that a switch between two windows initiates an association between them, and this increases as more switches are effected. This incremental nature of the data can be represented by a graph, $G_a(V, E)$ that evolves with the introduction of new nodes (documents) and edges (switches). The edge weights represent the association between the nodes.

The clustering of those documents that pertain to the same task can also be represented by a graph, $G_c(V, E)$ that will also need to be updated incrementally, since a switch to a new document will trigger a decision process do deal with the change. The changes to $G_c(V, E)$ that our clustering algorithm has to deal with include:

i. the *creation* of a new cluster: when the association between two nodes exceeds the $\epsilon$ threshold;
ii. *merging* of two clusters: when either a core or border node in one cluster gets strongly associated with another core or border node in another cluster;
iii. *absorption* (a growing cluster): when the association between a core or a border node and a new node exceeds the threshold $\epsilon$.

Consider a typical situation whereby the initial edge weight between two nodes $u$ and $v$ is $> \epsilon$. With increased switches, the association strength increases since the edge weight will decrease and possibly become $\leq \epsilon$. At this point, either, or both, of $u$ and $v$ can become core nodes (depending on $\eta$) with the consequence that nodes in their neighbourhood can either change status as well,

form a cluster or merge with an existing one. It might also be the case that either $u$ or $v$, or both, become border nodes, and thus form a potential cluster.

In **iDeTaCt** we use an association edge-weighting function $W : \mathbb{R} \rightarrow \mathbb{R}$ that maps the number of window-switches between two documents to an edge weight $w(u, v)$ in $G_a(V, E)$. We do not consider the direction of the edge, that is, an edge from node A to node B is considered the same as an edge from B to A. The resulting edge weight is inversely proportional to the number of window-switches. Thus a high number of switches will result in a lower edge weight. This is similar to the *proximity* and *influence* functions used in [7, 8] respectively, whereby two nodes are considered to be closer together if the edge weight between them is less.

We compute the number $n$ of edges between two nodes as a fraction of a defined maximum number of edges, $h$. This maximum number is empirically set to 10 which is considered to be sufficiently indicative of a strong association between any two documents. Thus if the number of edges is 1, the value passed on to the $W$ would be equal to $\frac{1}{10}$.

The edge-weighting function $W(\frac{n}{h})$ is based on the Epanechnikov kernel [8] and is defined as:

$$W(x) = \begin{cases} \frac{3}{4}(1 - x^2) & |x| \leq 1 \\ 0 & \text{else} \end{cases} \tag{1}$$

**iDeTaCt** takes as parameters the newly generated edge $e$ and the old clustered graph $G_c(V, E)$ and works as follows:

- The association edge-weighting function *computeAssociation* takes as parameter the number of switches between $u$ and $v$ and returns the updated weight $w(u, v)$ of $e$.
- This weight is used to increase the ranking of nodes $u$ and $v$ within a cluster. If $w(u, v)$ is less than or equal to $\epsilon$ the node's ranking is increased by a factor of 0.85, otherwise it is increased by a factor of 0.15. This is in line with the way that Firefox's frecency algorithm[7] assigns a bonus to recently viewed pages.
- If edge $e$ does not exist in $G_c$ then $e$ is added and $G_c(V, E)$ is updated. This results in endpoints $u$ and $v$ of $e$ becoming connected in $G_c(V, E)$.
- Then for both $u$ and $v$, if they are not core, we consider all their incident edges to check whether the changes have effected their status.
- If there are $\eta$ or more such edges incident on node $u$ then its status is set to *core*.
- If $u$ is not *core* but is adjacent to a *core* node then its status is defined as *border*.
- Nodes that are neither *core* nor *border* are considered as *noise* and are placed on a stack for later consideration.

[7] `https://developer.mozilla.org/en-US/docs/Mozilla/Tech/Places/Frecency_algorithm`

– Then **iDeTaCt** calls **updateClusters** which performs a Breadth-First-Search over $G_c$ to find the updated induced subgraphs. Each subgraph represents a cluster.
– In the process, **updateClusters** tries to include particular nodes from the stack that are still labelled as *noise* using the procedure *findWeakNodes*.
– In *findWeakNodes*, if a node $i$ is found to be a neighbour to $u$ and $v$ in $G_c(V, E)$ which have a status of core than the status of $i$ is changed to *weak* and it is added to $G_c(V, E)$. Although such nodes have a weak relation with the surrounding nodes, in that they fall short of the $\epsilon$ threshold, they are connected to nodes which in turn are strongly connected.
– **iDeTaCt** returns a list of clusters that can be visualised through the *PiMx-Viz* and the *PiMx-Clusters* components.

## 6 Evaluation

In our evaluation we wanted to verify whether **iDeTaCt** was able to: (i) identify those nodes that belong together in a task-cluster and (ii) identify when a switch between two nodes is effectively a task-switch.

For the evaluation we made use of the *PiMx* framework to simulate the task execution trails of the users from groups 1 (considered as the baseline group) and 2 (interleaved tasks). Details about the number of pages visited and the number of switches made by participants in these two groups can be seen in Fig. 2 and Fig. 3. We did not use the data from Group 3 since we wanted to initially evaluate our approach on data that was collected during a single session.

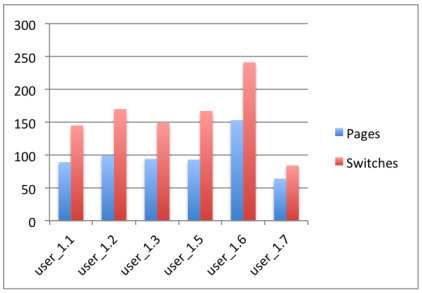


Fig. 2: Pages/Switches for Grp 1

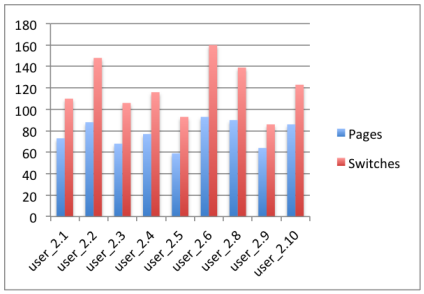


Fig. 3: Pages/Switches for Grp 2

For each trail we used **iDeTaCt** with core parameter $\eta$ values of 1, (*D_1*) and 2, (*D_2*). With $\eta = 1$, two nodes will become core when they are connected by a single edge whose weight $w(u, v) \leq \epsilon$. Similarly with $\eta = 2$, a node will need to be connected to two other nodes through two similar edges. $\epsilon$ was set to the maximum value of 0.72 which is equivalent to a minimum of two window switches (using equation Eq. 1).

We used standard information retrieval metrics to evaluate the clusters for each of the three tasks. These metrics involve (i) *precision*, defined as the percentage of documents correctly assigned to a task-cluster over the total number of documents in the task-cluster, (ii) *recall*, defined as the percentage of documents correctly assigned to a task-cluster over the total number of documents that should have been assigned to that task-cluster, and (iii) *F1-measure*, defined as the combined measure that assesses a trade off between *precision* and *recall*. Whenever the algorithms generated two or more clusters for documents from the same task, we considered the cluster which was more representative of the task, that is, it contained the highest number of documents. In the case of the interleaved tasks in Group 2, we expect the algorithm to be able to cluster the interleaved tasks as if they were actually none interleaved.

Precision was 100% when we tried both *D_1* and *D_2* on all the task-trails from Group 1. However when we used *D_1* on the dataset from Group 2 it was less then 100% in all cases except one. When we changed $\eta$ to 2 on the data from Group 2 we got 100% precision in 66% of the cases, in all the 3 tasks. In the rest, the precision was less than 100% for only one of the tasks.

We focus on the more interesting recall and F1-measure. The averaged results are shown in Fig. 4 and Fig. 5 respectively. We again compute the recall and F1-measure for all the task-trails from Groups 1 and 2, and we do this for every task separately.

Recall for the task-clusters generated on the data from Group 1 was highest when we used *D_1*, with the averaged recall being highest for task 2, at 70.7%. Task 3 had the lowest averaged recall at 36.5% due to the algorithm generating multiple, 2 or 3-node clusters. The possible reason for this could be due to the familiarity of the participants with this topic. The fact that a very important music event was forthcoming when the experiment was conducted, might have effected the participants' information seeking behaviour with many of them knowing where to search and thus the number of re-visits was low.

The F1-measure for the task-clusters from Group 1 was consistent with the recall and was again highest when we used *D_1*. As expected, the number of captured nodes with this value of $\eta$ was higher than with *D_2* and the resultant task-clusters where denser, yet still separate. Task 3 had once again the least averaged F1-measure irrespective of $\eta$.

We now consider the recall and F1-measure based on the data from Group 2. The values obtained for the task-clusters associated with task 2 were the highest, and ranged between 50% and 60%. This trend is in line with the results we got for the task-clusters from Group 1, however both values for task 3 are slightly higher then those for task 1 except for the F1-measure based on *D_1*.

The recall and F1-measure based on the data of 2 of the participants from this group resulted in exceptionally low values for the task-cluster related to task 3. This was independent of $\eta$. For another user from the same group, we observed the same low values for the task-cluster related to task 1. The common feature observed across the data of these 3 participants was that the relevant task was fragmented in multiple 2 or 3-node clusters.

From the generated graphs it was possible to observe that all the participants tend to open up a number of documents which they only visit once. This was more accentuated in almost 50% of the cases from Group 2. For some of these, we manually inspected their log file and found that in fact these participants typically opened up a number of tabs in succession, as in the case of a result page associated with some query. They however only visited some of those opened tabs once. Different individuals however did revisit some of the documents multiple times and these acted like hubs/authorities to the other accessed documents thus allowing for the generated clusters to be more consistent.

The fact that **iDeTaCt** managed to cleanly separate all the tasks, using the data of all the participants from the baseline group is already encouraging. It is even more encouraging when we used the interleaved tasks and found the algorithm to be sufficiently reliable in more than 50% of the cases, even though this was at the expense of capturing less documents from the referenced tasks.

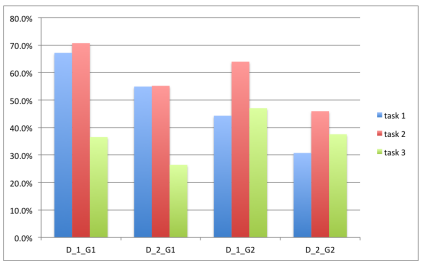


Fig. 4: Average Recall

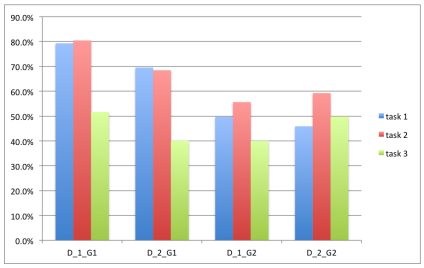


Fig. 5: Averaged F1-measure

## 7 Future Work and Conclusion

We intend to take into consideration the type of edge, which requires us to extend *iDeTaCt* to handle navigational events, especially those relating a query with its result pages. Navigational events accounted on average for 30% of all switches performed by each participant. We also plan to consider the window captions and apply a similarity function, such as cosine similarity, over this content, based on the clusters generated by *iDeTaCt*. Furthermore, we want to make use of *iDeTaCt* for task-resumption and in-line task/document recommendations. We intend to evaluate these extensions and compare our results.

In this paper we described the experiment we conducted to collect data for the evaluation of our *iDeTaCt* incremental graph clustering algorithm. We executed the algorithm over the task collections and used the *PiMx* framework to analyse its performance. The results showed a high precision and recall over the data from the baseline group and a recall of more than 50% over the data for the interleaved tasks. This is considered to be very encouraging and motivates us to further investigate how to improve our algorithm.

## References


1. Bernstein, M., Shrager, J., Winograd, T.: Taskpose: Exploring Fluid Boundaries in an Associative Window Visualization. In: 21st ACM Symposium on User Interface Software and Technology, pp. 231-234. ACM Press New York, NY, USA (2008)
2. Charikar, M., Chekuri, C., Feder, T. and Motwani, R.: Incremental clustering and dynamic information retrieval. In Proceedings of the twenty-ninth annual ACM symposium on Theory of computing (STOC '97), pp. 626-635. ACM, New York, NY, USA (1997)
3. Costache, S., Gaugaz, J., Ioannou, E., Niederee, C., Nejdl, W.: Detecting contexts on the desktop using bayesian networks. In: Desktop Search Workshop co-located with SIGIR (2010)
4. Dabbish L., Mark, G. and Gonzlez, V.M.: Why do I keep interrupting myself?: environment, habit and self-interruption. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11), pp.3127-3130. ACM, New York, NY, USA (2011)
5. Dumais, S., Cutrell, E., Cadiz, J.J., Jancke, G., Sarin, R., and Robbins, D.C.: Stuff I've seen: a system for personal information retrieval and re-use. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR '03), pp.72-79. ACM, New York, NY, USA (2003)
6. Ester, M., Kriegel, H.-P., Sander, J., Wimmer, M., and Xu, X.: Incremental clustering for mining in a data warehouse environment. In Proc. of 24th VLDB Conference (1998).
7. Falkowski, T., Barth, A. and Spiliopoulou, M.: DENGRAPH: A Density-based Community Detection Algorithm. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI '07). IEEE Computer Society, Washington, DC, USA, 112-115 (2007)
8. G*ü*nnemann, S. and Seidl, T.: Subgraph Mining on Directed and Weighted Graphs. In Proc. of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2010), pp. 133-146. Hyderabad, India. Springer - Heidelberg, Germany. (2010)
9. Jones, W.P., Teevan, J.: Personal Information Management. ISBN 9780295987378, University of Washington Press (2007)
10. Mayer, M.: Visualizing web sessions: improving web browser history by a better understanding of web page revisitation and a new session- and task-based, visual web history approach. PhD thesis, University of Hamburg (2008)
11. Morris, D., Ringel Morris, M., Venolia, G.: Searchbar: a search-centric web history for task resumption and information re-finding. In: 26th annual SIGCHI conference on Human factors in computing systems, CHI '08, pp. 1207-1216. ACM Press, New York, NY, USA (2008)
12. Oliver, N., Smith, G., Surendran, A.C.: SWISH: Semantic Analysis of Window Titles and Switching History. In: 10th International Conference on Intelligent User Interfaces, pp. 194-201. ACM Press, New York, NY, USA (2006)
13. Wang, Q. and Chang, H.:Multitasking bar: prototype and evaluation of introducing the task concept into a browser. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10). ACM, New York, NY, USA, 103-112. (2010)

# Creativity Profiling Server: Modelling the Principal Components of Human Creativity over Texts

George Panagopoulos, Pythagoras Karampiperis, Antonis Koukourikos, Sotiris Konstantinidis

Computational Systems & Human Mind Research Unit, Institute of Informatics & Telecommunications, National Center for Scientific Research "Demokritos", Greece

{gpanagopoulos,pythk,kukurik,skonstan}@iit.demokritos.gr

**Abstract.** Within the field of Computational Creativity, significant effort has been devoted towards identifying variegating aspects of the creative process and constructing appropriate metrics for determining the degree that an artefact exhibits creativity with respect to these aspects. However, the formalization of a person's creativity (i.e. a creativity user profile) as a derivative of such creations is not straightforward, as it requires a transition to a space reflecting the core principles of creativity as perceived by humans. This becomes a necessity in domains where personalization goes beyond timely and personalized knowledge provision, targeting the encouragement and fostering of creative thinking. Thus, it becomes essential to develop methodologies for modelling creativity to support personalization based on creativity aspects / characteristics of users. The paper proposes a user modelling framework for formulating creativity user profiles based on an individual's creations, by transitioning from traditional computational creativity metrics to a space that adheres to the principal components of human creativity. Furthermore, the paper presents the Creativity Profiling Server (CPS), a system implementing the aforementioned user modelling framework for computing and maintaining creativity profiles and showcases the results of experiments over storytelling educational activities.

**Keywords:** Human Creativity Modelling, Creativity Profiling, Computational Creativity

## 1 Introduction

Human creativity is a multifaceted, vague concept, combining undisclosed or paradoxical characteristics. As a general notion, creativity adheres to the ability to move beyond traditional and established patterns and associations, by transforming them to new ideas and concepts or using them in innovative, unprecedented contexts and settings [1]. The usage of computational methods for producing creative artefacts, as well as, unveiling the essence of human creativity and using computers understanding it, is the subject of extensive debate [2]. Along with such philosophical approaches, research results from neuroscience should also be considered in the process of reveal-

ing/ understanding the human creative process. In general, the creativity of a person can be expressed qualitatively by taking into account its origin in psychometric or cognitive aspects of their thinking process [3]. An example of the former is the work of [4], who examine how the human mind perceives complex auditory stimuli e.g. music. In this case, they look at the brains of improvising musicians and study what parts of the brain are involved in the kind of deep creativity that happens when a musician is really in the groove. Their research has deep implications for the understanding of creativity of all kinds. In any case, while machines can mimic human creativity, or provide the necessary stimuli for encouraging and promoting the production of creative ideas and artefacts, it is not straightforward to assess the exhibited creativity by using automated techniques. Rather, most efforts have been focused on analyzing creativity on different aspects and producing different metrics, based on the nature of the examined artefacts.

Hence, the core assumption for building a user's creativity profile, is that his/her creativity is showcased by his/her creations, named Creativity Exhibits. These exhibits can follow different modalities, corresponding to the aforementioned reasoning patterns, e.g. texts, diagrams/pictures, actions etc.

The calculation of a creativity profile, constitutes the process of (a) measuring the creativity expressed by given creativity artifacts; (b) associating these measurements with dimensions of human creativity corresponding to the given dimension.

For achieving (a), we employ creativity metrics derived from computational creativity and formulate them in accordance to the characteristics of the examined exhibits. A number of different creativity metrics are proposed from the literature on computational creativity.

More specifically, Novelty reflects the deviation from existing knowledge/ experience and can be measured as a difference metric between what is already known and the given piece of content. Novelty is a generally accepted dimension of creativity within the area of computational creativity and an essential candidate for measuring elements of creativity within the human-created content when interacting with the machine. It has been used as a heuristic for driving the generation of novel artefacts in exploratory creativity [3] known as novelty search, an approach to open-ended evolution in artificial life [5]. Surprise is another essential characteristic which may be represented as the deviation from the expected [6]. The higher the deviation the higher the perceived surprise. Surprise offers a temporal dimension to unexpectedness [7]. Likewise, impressive artefacts readily exhibit (ease of recognition) significant design effort and may be described via two heuristics, Rarity (rare combination of properties) and Recreational Effort (difficult to achieve) [8]. These four metrics will be used to construct the creativity profile of a human user, as expressed by the artefacts that this user has been constructed alone or as a participating member of a group of users. In the case of Textual Exhibits, examples of such artefacts include a written story, a dialogue and any other textual creation.

In our previous work [9] we presented the formulization of the Computational Creativity Metrics for Novelty, Surprise, Rarity and Recreational Effort over textual artefacts. In the present work, we use these text-based metrics for the core aspects of creativity and examine their conformance with the human perception of what constitutes a

creative artefact. We proceed to identify the deviations between these two perspectives (computational metrics and human judgment) and propose a model for transforming the automatic measures to a space that more accurately reflects the human opinion. In this way, the constructed human creativity profiles can be used for providing personalized material / content that is suitable for a specific user or addresses his/her limitations regarding creativity.

The rest of the paper is structured as follows. We examine the correlation of the proposed metrics with the human perception of creativity. Afterwards, we build on these observations to propose a transition model from computational metrics to a two-dimensional orthogonal space which aims to closely reflect the way human beings perceive creativity. We present the experiments for assessing the effectiveness of the proposed model towards this goal, describe the architecture and functionality of the Creativity Profiling Server, a system that incorporates the proposed model and report on the experiments for a preliminary evaluation of the system. Finally, we summarize the present research and report on our next steps.

## 2 Correlation of Computational Creativity Metrics With the Human Perception of Creativity

In order to assess the adherence of the proposed metric formulization with the human perception for creativity, we organized and conducted an experimental session based on storytelling activities. For the execution of the experiment, we employed forty (40) human participants, split in ten (10) teams of four (4) members each. All teams were asked to construct a story, on a specified premise, the survival of a village's habitants under a ravaging snow storm. The stories were created incrementally, with twenty (20) fragments produced for each story.

Following the completion of the stories, the teams were organized in two groups, each consisting of five teams. Without any interaction between the groups, each team was called to rate the stories of the remaining four teams belonging to their group, using a rank-based 4-star scale (i.e. the best story received 4 stars, the second-best story received 3 stars etc.). In this way, we obtained a ranked list of the five stories in each group. The goal of our experiment was to determine if, using the ranked lists of one of the test groups and a formalized representation of the computational creativity metrics, we can identify their correlation and examine if the distribution of values for the metrics follow the pattern of human judgment. To this end, we define a constrained optimization problem over functions of the aforementioned metrics, which is described below.

### 2.1 Extracting a Model for the Human Perception of Creativity

Each artefact (story) $S_n$ is characterized (via the application of the computational creativity metrics presented in the previous section) [9] by a set of 4 independent properties $g^{S_n} = ({g_1}^{S_n}, {g_2}^{S_n}, {g_3}^{S_n}, {g_4}^{S_n})$ where $g_1$ stands for "Novelty", $g_2$ for "Surprise", $g_3$ for "Rarity" and $g_4$ for "Recreational Effort". We define as partial creativi-

ty function (PCF) related to artefact property $g_k$ a function that indicates how important is a specific value of the property $g_k$ when calculating the creativity of an artefact $S_n$. This function is defined by the following formula:

$PCF_{g_k}({g_k}^{S_n}) = w_{g_k} * \left( \frac{c_{g_k} * \left(1 - d_{g_k}\right)}{e^{\left( a_{g_k} * g_k {}^{S_n} + b_{g_k} \right)^2}} + \frac{d_{g_k}}{2} \right)$, where ${g_k}^{S_n} \in [0,2]$is the value of property $g_k$ for the artefact $S_n$, and $0 \leq a_{g_k} \leq 5,\ -4 \leq b_{g_k} \leq 4,\ \ 0 \leq c_{g_k} \leq 1,\ 0 \leq d_{g_k} \leq 2$ are parameters that define the form of the partial creativity function, whereas $0 \leq w_{g_k} \leq 1$ represents the weight of property $g_k$ in the calculation of the overall creativity. The calculation of the above parameters for all $g_k$ properties lead to the calculation of the complete creativity function (CCF), as the aggregation of the partial creativity functions, as follows: $CCF(g^{S_n}) = \frac{1}{4} * \sum_{k=1}^{4} PCF_{g_k}({g_k}^{S_n})$

If $CCF_{S_1}$ is the complete creativity of an artefact $S_1$ and $CCF_{S_2}$ is the complete creativity of an artefact $S_2$, then the following properties generally hold for the complete creativity function:

$$CCF_{S_1} > CCF_{S_2} \Leftrightarrow (S_1)P(S_2)$$

$$CCF_{S_1} = CCF_{S_2} \Leftrightarrow (S_1)I(S_2)$$

where P is a strict preference relation and I is an indifference relation, as perceived by humans when evaluating the creativity of these artefacts.

Given a preference ranking of a reference set of artefacts, we define the creativity differences $\Delta = \left(\Delta_1, \Delta_2, \dots, \Delta_{q-1}\right)$, where q is the number of artefacts in the reference set and $\Delta_i = CCF_{S_i} - CCF_{S_{i+1}} \geq 0$ is the creativity difference between two subsequent artefacts in the ranked set.

We then define an error parameter $E$ for each creativity difference:

$$\Delta_i = CCF_{S_i} - CCF_{S_{i+1}} + E_i \geq 0$$

We can then solve the following constrained optimization problem:

$$Minimise \sum_{i=1}^{q-1} (E_i)^2 \ s.t. \begin{cases} \Delta_i \geq 0, if\ (S_i)P(S_{i+1}) \\ \Delta_i = 0, if\ (S_i)I(S_{i+1}) \end{cases}$$

This optimization problem leads to the calculation of the partial creativity function parameters for each property $g_k$.Based on these values and the human assessment of the story rankings, the results of the constrained optimization problem defined in the previous section resolves in the calculation of the partial creativity parameters (a, b, c, d and w). Regarding the impact of the various metrics in the computation of the overall creativity, we observed that Novelty is generally considered a particularly positive attribute creativity-wise for the stories, its partial creativity (PC) increasing as its value increases (see Figure 1). In contrast, the remaining metrics reached their maximum partial creativity at a certain value, after which their partial creativity started to decrease, indicating that e.g. recreational effort greater than a certain point is not perceived as a direct indication of creativity (see Figure 1).

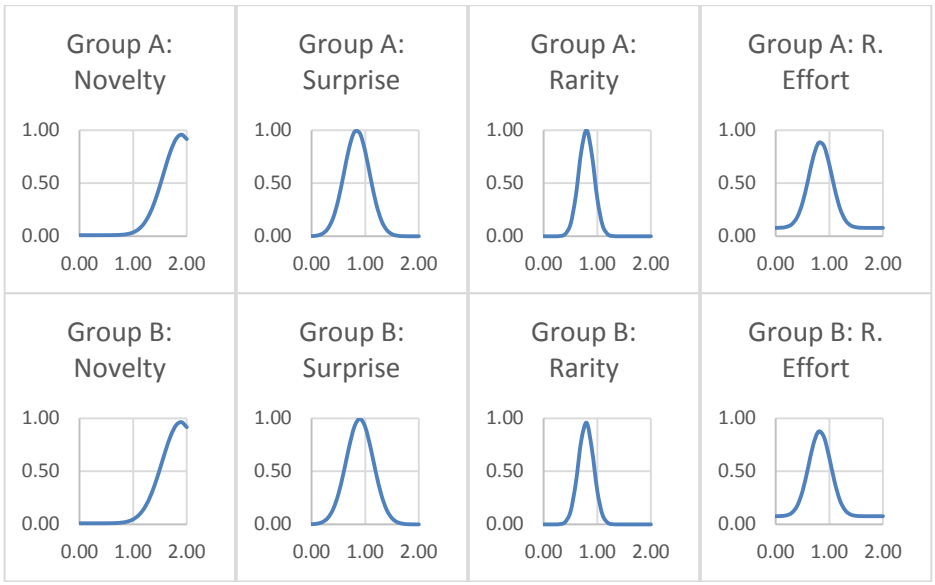


**Fig. 1.** PCs of Computational Creativity Metrics wrt their value (Group A & B respectively)

Hence, the obtained results indicate that, while the proposed computational creativity metrics are correlated with the perception of humans for creativity, this correlation is not direct for all metrics. The following section discusses on the implications of these observations and details our approach for using the proposed metrics towards building a dimensional plane that more accurately reflects the human perspective for creativity.

## 3 Transferring Computational Creativity Metrics to the Human Perspective

As stated, each textual artefact can be described by 4 computational creativity metrics, namely, Novelty, Surprise, Rarity and Recreational Effort. Following the formulation of the creativity metrics, therefore, the next hypothesis that was examined was the reduction of the dimensional space for representing creativity as expressed through creative artefacts, in an orthogonal space. In order to effectively conceptualize human creativity, orthogonality is a particularly desirable attribute of the conceptualization space to be used, since it allows the examination of independent variables when trying to analyse and influence / encourage certain creativity aspects. Hence, the first step towards identifying the adherence of the computational creativity metrics with the human perspective is to examine the orthogonality of the proposed metrics formulation. To this end, we ran an experiment for calculating the four basic computational creativity metrics on two datasets derived from distinct and distant domains, and determined whether the four metrics are orthogonal.

The first dataset comprised transcriptions of European Parliament Proceedings [10]. Given the formulation of computational creativity metrics described in [9], we

consider as a "story" the proceedings of a distinct Parliament session and as a fragment the speech of an individual MP within the examined session. The second dataset was derived from a literary work, Stories from Northern Myths, by E.K. Baker, available via the Project Gutenberg collection. In this case, the story is a book chapter and the story fragment is a paragraph within the chapter.

**Table 1.** Computational Metrics Correlation: Formal Verbal Transcriptions

| | Novelty | Surprise | Rarity | R. Effort |
|---|---|---|---|---|
| Novelty | 1.00000 | 0.13393 | 0.12329 | -0.40681 |
| Surprise | 0.13393 | 1.00000 | 0.26453 | -0.43151 |
| Rarity | 0.12329 | 0.26453 | 1.00000 | -0.33499 |
| R. Effort | -0.40681 | -0.43151 | -0.33499 | 1.00000 |

**Table 2.** Computational Metrics Correlation: Literary Work

| | Novelty | Surprise | Rarity | R. Effort |
|---|---|---|---|---|
| Novelty | 1.00000 | -0.64243 | 0.10392 | -0.10762 |
| Surprise | -0.64243 | 1.00000 | 0.07376 | -0.02538 |
| Rarity | 0.10392 | 0.07376 | 1.00000 | -0.03882 |
| R. Effort | -0.10762 | -0.02538 | -0.03882 | 1.00000 |

In total, we examined 50 distinct parliament sessions from the Europarl dataset and 40 chapters from the storybook. Based on the obtained results, we calculated the correlation between the four computational creativity metrics. Tables 1 and 2 provide the correlation values between the four metrics. It is evident that the computational creativity metrics by themselves are not orthogonal. In order to better approximate the human perception for creativity, we propose the following abstraction for modelling the examined aspects of creativity to a space more closely resembling human thinking:

*Novelty* is the perspective to be held as the one dimension of the dimensional space, as the conducted showed that it has a monotonic incremental relation with the perception of humans on what is creative. Further more, it is a generally accepted dimension of creativity. [11]

*Atypicality*, that is, the tendency to deviate from the norm without actually breaking through. In other words, to what extend (without necessarily being novel) the artefact differs from the ordinary (thus being surprising, rare and difficult to construct)

We consider Atypicality as a combination of the Surprise, Rarity and Recreational Effort metrics, each bearing a different weight towards determining Atypicality. These two axes also provide a rough conceptualization of the two major qualitative aspects of creative work: whether the said work is visionary, i.e. it provides a groundbreaking approach on a given field; and whether it is constructive, i.e. it uses in a novel way established techniques and ideas in order to produce a high-quality artefact. As stated, Novelty has an analogous and close to monotonic association with the human judgment for creativity. Therefore, and in order to satisfy our requirement of

orthogonality, we consider Novelty as the strictly defined dimension of our space and seek for the formulation of Atypicality that results to a dimension orthogonal to Novelty.

More specifically, let Atypicality of a text $t$ be the normalized weighted sum of its Surprise, Rarity, and Recreational Effort: $A(t) = \frac{w_s Sur(t)+w_r Rar(t)+w_e Eff(t)}{w_s+w_r+w_e}$, with $w_s, w_r, w_e \in [-1,1]$. We aim to find the weight values that constitute Atypicality orthogonal to Novelty, i.e. those weight values for which Correl(Novelty, Atypicality)= 0. We thus define the following optimization problem:

$$Minimise \sum_{i=1}^{n} (\ Correl(Novelty_i, Atypicality_i)\ )^2, \text{ s.t. } w_s, w_r, w_e \in [-1,1]$$

where n is the number of the combined datasets.

Although the search space of the optimization problem above is highly non-linear solving this problem is straightforward. The resulting model defines two orthogonal axes, Novelty and Atypicality, which define the space for measuring and characterizing the observed creativity, as an Euclidean vector, the length of which indicates the quantitative aspect of the creativity exhibited by the artefact, while its direction indicates the tendency for either Novelty (visionary creativity) or Atypicality (constructive creativity). The following tables present the novelty and atypicality in the two datasets, as well as, the correlation between these two dimensions for the found optimum weight values.

**Table 3.** Correlation of Creativity Dimensions: Formal Verbal Transcription

| | Novelty | Atypicality |
|---|---|---|
| Novelty | 1.00000 | 2.986E-07 |
| Atypicality | 2.986E-07 | 1.00000 |

**Table 4.**Correlation of Creativity Dimensions: Literary Work

| | Novelty | Atypicality |
|---|---|---|
| Novelty | 1.00000 | 1.436E-07 |
| Atypicality | 1.436E-07 | 1.00000 |

## 4 The Creativity Profiling Server

The Creativity Profiling Server (CPS) allows the storage, maintenance and update of creativity profiles of users using creativity exhibits that are produced from applications of the outside world. CPS provides a simple and straightforward API in order to expose its functionalities and to facilitate the communication with the outside world. Through the CPS API, the example application can submit creativity exhibits and receive the corresponding creativity measurements, create group of users and finally receive the creativity profile of a user. The aforementioned functionalities and the internal structure of CPS are depicted in Figure 2.

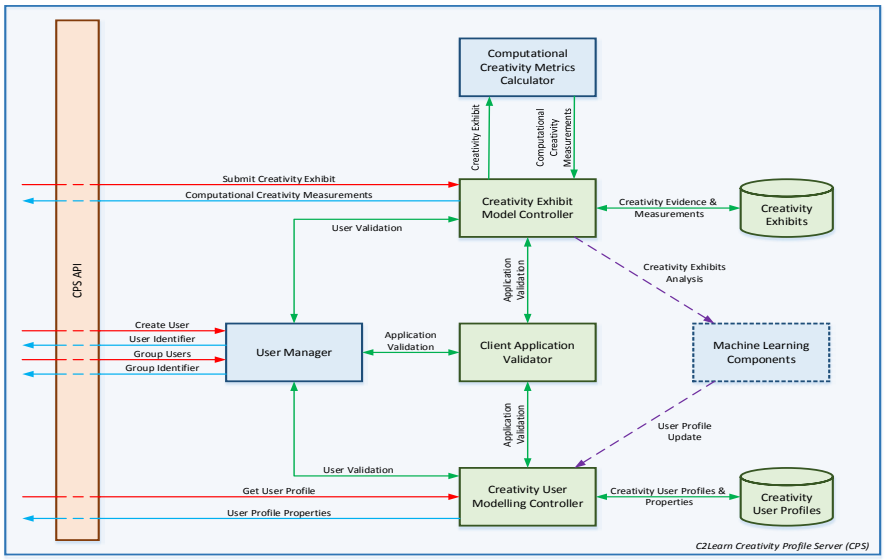


**Fig. 2.** CPS Architecture

The distinct modules incorporated in the CPS Architecture are the following:

- Client Application Validator: The module is responsible for ensuring that a client request is originated from an application registered to CPS.
- User Manager: This module is responsible for ensuring that client requests contain a valid user profile ID. Also User Manager is responsible for creating and destroying groups by joining and disjoining user profile properties respectively
- Creativity Exhibit Model Controller: This module is responsible for storing, maintaining and updating the creativity exhibits delivered by applications and also forward the creativity exhibits to the Computational Creativity Metrics Calculator: This module is responsible for calculating all the metrics of a creativity exhibit regarding of its type.
- Creativity User Modelling Controller: This module is responsible for storing, maintaining and updating the Profile Properties of each User Profile in CPS. Also this module delivers to client applications the properties of particular user profiles.
- Machine Learning Components: This module is responsible for calculating the value of the Creativity Profile Properties of a given user.

In a typical situation an application creates a user through the CPS API. The CPS API send the request to the User Management. Afterward User Management verifies through the Application Validation module that the application is registered to CPS. Since the application is validated User Management creates a unique user profile id and sends it to the application. Since a user profile is created then the application can submit creativity exhibits of this user. More specifically the application submits the creativity exhibit to the CPS API along with type of the creativity exhibit and the timestamp the creativity exhibit was created. After submission the API sends the crea-

tivity exhibit and its type to the Creativity Exhibit Model Controller module. After validating the user and the application through the User Management and the Application Validator respectively, the module sends the creativity exhibit to the Computational Creativity Metrics Calculator module. The Computational Creativity Metrics Calculator returns back the measurements of the creativity exhibit. Afterwards, the Creativity Exhibit Model Controller module stores the creativity exhibit along with the measurements to the CPS database. Finally, the Creativity Exhibit Model Controller invokes the Machine Learning Components. Machine Learning Components take as input the creativity exhibit and calculate the values of the profile properties of the user. Afterwards the newly calculated values are send to the Creativity User Modelling Controller module, which stores the values to the CPS database.

Once a user creativity profile is created, then the application can request through the CPS API the User Profile Properties and also the Model which describes the profile. After sending the request to the API, the request is redirected to the Creativity User Modelling Controller module. This module, after validating the user and the application through the User Management and the Application Validator respectively, retrieves from the CPS database the properties for the corresponding user and send them back to the application.

## 5 Incorporation of the model in CPS

Following the definition of the model, we combine within CPS the Surprise, Rarity and Recreational Effort metrics in order to form another metric, which we call Atypicality and is orthogonal to Novelty. Atypicality is calculated as a weighted average of Surprise, Rarity and Recreational Effort, as follows: $AT_i = \frac{w_S * S_i + w_R * R_i + w_E * E_i}{w_S + w_R + w_E}$, where: i refers to an artifact, Si , Ri ,Ei and ATi to Surprise, Rarity, Recreational Effort and Atypicality metrics respectively for the given artifact i, and wS,wR and wE are positive weights assigned to Surprise, Rarity and Effort respectively, in order to calculate the Atypicality metric in a way as much uncorrelated (and thus, orthogonal) with Novelty as possible. A user's Creativity Profile, thus, consists of a two-dimensional vector expressing two types of user's creativity. The Visionary Creativity, which is measured by the Novelty metric, and the Constructive Creativity, which is measured by the Atypicality metric. CPS gathers all Creativity Exhibits (artefacts) that are produced by its users within external applications. In discrete time intervals, which we call Time Window, CPS calculates and/or updates the Creativity Profile of each user. The calculation of the creativity profiles for the users of the CPS is a repeated (once per Time Window) two-phase process, and is explained below:

Phase A: *Calculation of optimum Computational Creativity Metric Weights for the Application Domain*

We aim to find/ update the weight values [wS,wR,wE] of Surprise, Rarity and Recreational Effort that constitute Atypicality orthogonal to Novelty, i.e. those weight values for which Correl(N, AT) = 0. The optimum vector [wS,wR,wE] will be used in

Phase B for the calculation of the users' Creativity Profiles for the new CPS Time Window.

We thus define the following non-linear optimization problem:

$$Min.\ Correl(N, AT)^2, st.\ w_S, w_R, w_E \geq 0\ , w_S + w_R + w_E \neq 0$$

Each time where a new CPS Time Window starts, we solve the above minimization problem for all the artefacts of the application domain (all the creativity exhibits collected for all CPS users and for all CPS Time Windows so far). It is evident that in each execution of this process there is a strong probability of discovering a new vector [wS,wR,wE] that makes Atypicality (AT) more orthogonal to Novelty (N). In order to reduce the sensitivity of the system to this continuous change, we update the vector [wS,wR,wE] to be used in Phase B with the new vector retrieved by solving the optimization problem defined in Eq. 1 only when the improvement (minimization) in Correl(N, AT)2 exceeds 5%.

Phase B: *Construct/update of Users' Creativity Profiles*

A user's creativity profile is determined by the creativity exhibits produced by the user alone or as a member of a group. Groups are treated by CPS as a user, meaning that CPS will construct a creativity profile also for each group. In this case, the creativity profile is constructed/ updated based on the creativity exhibits of the group during the last (just finished) time window. In the case of simple users (not groups) their creativity profile is constructed/ updated based on all the creativity exhibits they constructed (either alone or as a group member). The first step for computing the creativity profiles is to transform the space (N,S,E,R) to the space (N, AT) and compute the average of N and AT measures for the creativity exhibits for a given user and for the time window that just finished, as follows:

*B1. Calculate Average Novelty and Atypicality of Creativity Exhibits*

In the general case, let a user U which participates in groups UG. In the case of computing the creativity profile of a group, we have only the user U, which represents the group. Such a user cannot be part of other groups. Let $E_T \equiv [\overline{Novelty}, \overline{Atypicality}]$ of a user U, calculated for the creativity exhibits in the time window T, after the transformation of the space (N,S,E,R) to the space (N, AT) using the optimal weight vector [wS,wR,wE] (calculated in Phase A). Let also $G_T \equiv [\overline{Novelty}, \overline{Atypicality}]$ of a user U, calculated for the creativity exhibits of UG in the time window T, after the transformation of the space (N,S,E,R) to the space (N, AT) using the optimal weight vector [wS,wR,wE] (calculated in Phase A).

The overall Average Novelty and Atypicality (PT) of all creativity exhibits for user U is calculated as a fusion of ET and GT, relying on the analogy of the user's and the groups' achievements. If the user's creativity (ET) surpasses the creativity exhibited within his/her participation in groups (GT), then only ET is considered. Otherwise, a part of the difference between groups' creativity and user's creativity is also considered, as follows:

$$P_T = \begin{cases} E_T & E_T \geq G_T \\ E_T + k * (G_T - E_T) & E_T < G_T \end{cases}, \text{ with } k = \frac{1}{2} + \frac{1}{2} * tanh(2 * [(G_T - E_T) - 1])$$

*B2. Calculate Visionary and Constructive Creativity of User*

Though all exhibits must be taken into account, the recent ones are considered more important, as they depict the exact current status of the user's creativity whereas past exhibits play a less vital role. To give our model an essence of decay through time, we use this formula: $C_T = \frac{P_T}{D} + \frac{D-1}{D} * C_{T-1}$, where: $C_T$ is the vector describing the Creativity of the user (or group) at the time window T, and $C_{T-1}$ at the time window T-1 respectively $C_T \equiv [Visionary\ Creativity, Constructive\ Creativity]$ and D, a proportional constant of decaying analogous to the timespan.

## 6 Preliminary CPS Evaluation

In order to obtain a preliminary assessment for the effectiveness of the proposed approach, we conducted a two-phase experiment in order to determine (a) the degree to which the selected computational creativity metrics conform to the opinion of experts regarding the creativity exhibited in a textual artefact and (b) the degree to which the proposed model for human creativity reflects the opinion of such experts.

For the purposes of the experiment, we employed twenty students who were asked to produce five stories each under pre-defined topics. For the first stage of the experiment, we sampled the stories produced during the aforementioned story writing session, randomly selecting two stories by each student, and asked five experts to rank them with respect to their creativity, as the latter is perceived by each of these experts. We then compared the ranking results with the ranking derived from the results produced by the CPS. For the second stage of the experiment, we picked the complete set of stories (i.e. five stories) for five of the users and asked from the same five experts to rank these users with respect to their creativity, using as evidence the produced stories. We then compared the expert ranking to the one produced by the CPS.

In order to evaluate the similarity between the rankings of the experts and the rankings of the CPS, for the textual exhibits' and the users' ranks, we employed a metric based on Kendall's Tau, defined by the following equation: $Success = \frac{1}{2} + \frac{N_{concordant} - N_{discordant}}{n(n-1)}$, where $N_{concordant}$ stands for the concordant pairs of ranked exhibits or users, $N_{discordant}$ stands for the discordant pairs when comparing the ordering of the experts and the CPS and $n$ is the number of the examined exhibits or the users. We calculated this metric for the series of textual exhibits rankings and the series of participating users rankings. The following table presents the summary statistics of the two Success metric series we had as an outcome.

**Table 5.** Correlation Coefficient between Expert and CPS rankings

| | Textual Exhibits | Users |
|---|---|---|
| Min Success | 0.58 | 0.56 |
| Average Success | 0.74 | 0.71 |
| Max Success | 0.89 | 0.88 |

## 7 Conclusions & Future Work

The work described in the present paper showcases our findings towards transitioning from computational creativity metrics associating specific attributes of text artefacts with creativity aspects to a creativity calculation model that better reflects the human perception of creativity. Furthermore, the present manuscript provides a summary of the architectural design and functionality of the Creativity Profiling Server (CPS).

Towards the continuation of our research, we aim to examine the effectiveness of the model in more complex experiments, examining textual exhibits from different domains and modalities (prose, poetry, speech) in order to obtain a more general reflection of the human perception of creativity. Observation over more open-ended experiments will likely lead to further refinements and extensions of the proposed human creativity model.

## 8 References


1. Zhu, X., Xu , Z., Khot, T.: How creative is your writing? a linguistic creativity measure from computer science and cognitive psychology perspectives. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, Association for Computational Linguistics, pp. 87--93, (2009).
2. Lubart, T.:How can computers be partners in the creative process: classification and commentary on the special issue. In *International Journal of Human-Computer Studies*, vol. 63, pp. 365--369, (2005).
3. Boden, M.: The Creative Mind, 2nd Edition. London: Routledge, (2004).
4. Limb, C. J., Braun, A. R.:Neural substrates of spontaneous musical performance: An fMRI study of jazz improvisation. In *PLoS One*, vol. 3, (2008).
5. Lehman, J., Stanley, K. O.:Exploiting Open-Endedness to Solve Problems Through the Search for Novelty. In *ALIFE*, pp. 329--336, (2008).
6. Macedo, L., Reisenzein, R., & Cardoso, A.: Modeling forms of surprise in artificial agents: Empirical and theoretical study of surprise functions. In *Proceedings of the 26th annual conference of the cognitive science society*, NJ: Erlbaum, pp. 873--878, (2004) .
7. Maher, M. L., Brady, K., Fisher, D. H.:Computational models of surprise in evaluating creative design. In *Proceedings of the Fourth International Conference on Computational Creativity*, (2013).
8. Lehman, J., Stanley, K. O.:Beyond Open-endedness: Quantifying Impressiveness. In *Artificial Life*, pp. 75--82, (2012).
9. Karampiperis, P., Koukourikos, A., Panagopoulos, G.:From Computational Creativity Metrics to the Principal Components of Human Creativity. In *Proc. of the 9th International Conference on Knowledge, Information and Creativity Support Systems*, Limassol, Cyprus, (2014).
10. Koehn, P.:Europarl: A parallel corpus for statistical machine translation. In *MT summit*, pp. 79--86, (2005).
11. G. Ritchie.:Some empirical criteria for attributing creativity to a computer program. In *Minds and Machines*, vol. 17, pp. 67--99, (2007).
12. Kövecses, Z.:A new look at metaphorical creativity in cognitive linguistics. In *Cognitive Linguistics*, vol. 21, pp. 663--697, (2010).
13. Veale, T.:An analogy-oriented type hierarchy for linguistic creativity. In *Knowledge-Based Systems*, vol. 19, pp. 471--479, (2006).

14. Lehrer, A.:Brendalicious. Lexical creativity, texts and contexts. Amsterdam, John Benjamin, , pp. 115-136, (2007).
15. Chiru, C. G.:Creativity Detection in Texts. In ICIW 2013, The Eighth International Conference on Internet and Web Applications and Services, (2013).

# Bitcoin Spread Prediction Using Social And Web Search Media

Martina Matta, Ilaria Lunesu, Michele Marchesi

Università degli Studi di Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
{martina.matta,ilaria.lunesu,michele}@diee.unica.it

**Abstract.** In the last decade, Web 2.0 services such as blogs, tweets, forums, chats, email etc. have been widely used as communication media, with very good results. Sharing knowledge is an important part of learning and enhancing skills. Furthermore, emotions may affect decision-making and individual behavior. Bitcoin, a decentralized electronic currency system, represents a radical change in financial systems, attracting a large number of users and a lot of media attention. In this work, we investigated if the spread of the Bitcoin's price is related to the volumes of tweets or Web Search media results. We compared trends of price with Google Trends data, volume of tweets and particularly with those that express a positive sentiment. We found significant cross correlation values, especially between Bitcoin price and Google Trends data, arguing our initial idea based on studies about trends in stock and goods market.

**Keywords:** Bitcoin, Twitter, Google Trends, Sentiment Analysis

## 1 Introduction

Bitcoin, a decentralized electronic currency system, represents a radical change in financial systems after its creation in 2008 by Satoshi Nakamoto [22]. Bitcoin stands for an IT innovation based on the advancement in peer-to-peer networks [21] and cryptographic protocols. Due to its properties, Bitcoin is not managed by any governments or bank. Like any other currency, a peculiarity of Bitcoin is to facilitate transactions of services and goods [20], attracting a large number of users and a lot of media attention.

Nowadays, Web 2.0 services such as blogs, tweets, forums, chats, email etc. are widely used as communication media, with satisfying results. Sharing knowledge is an important part of learning and enhancing skills. Through the use of social media services, team members have the opportunity to acquire more detailed information about their peers' expertise [7]. Social media data represents a collective indicator of thoughts and ideas regarding every aspect of the world. It has been possible to assist to deep changes in habits of people in the use of social media and social network. Twitter[10], an online social networking website and microblogging service, has become an important tool for businesses and individuals to communicate and share information with a rapid growth and

significant adoption. In addition, Twitter has rapidly grown as a mean to share ideas and thoughts on investing decisions.

In this work we analyze whether social media activity or information extracted by web search media could be helpful and used by investment professionals. There are several works that present predictive relationships between social media and bitcoin price where the relative effects of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships, are analyzed using cross-correlation such as [17] or linear regression analysis such as [6] or [5]. Social factors, that are composed of interactions among the actors of the market, may strongly drive dynamics of Bitcoin's economy [3].

We decided to apply automated Sentiment Analysis on shared short messages of users on Twitter in order to automatically analyze people's opinions, sentiments, evaluations and attitudes. We investigated whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict the Bitcoin market. We tried to discover if the chatter of the community can be used to make qualitative predictions about Bitcoin market, attempting to establish whether there is any correlation between tweet's sentiment and the Bitcoin's price [1]. The results suggest that a significant relationship with future Bitcoin's price and volume of tweets exists on a daily level. We also used Google Trends to analyze Bitcoin's popularity under the perspective of Web search, which provides a time series index of the volume of queries made by users in Google Search. We found a striking correlation between Bitcoin's price spread and changes in query volumes for the "*Bitcoin*" search term.

The body of this paper is organized in five major sections. Section 2, describes the background, section 3 presents the research steps of our study and section 4 summarizes and discusses our results. Finally, section 5 presents conclusions and suggestions for future work.

## 2 Background

In these decades, social web has been commercially exploited for goals such as automatically extracting customer opinions about products or brands, to find which aspects are liked and which are disliked [9]. In their work, Ye and Wu demonstrate how particularly interesting is the influence of Twitter users and the propagation of the information related to their tweets[18].

According to Alexa [12], Twitter had become the world's seventh most popular website by March 2015. Twitter [10] is an online social networking website and microblogging service that allows users to post and read text-based messages of up to 140 characters, known as "tweets". Launched in July of 2006 by Jack Dorsey, Twitter is now in the top 10 most visited internet sites with a total amount of 645,750,000 registered users. Java et al. affirm that it seems to be used to share information and to describe minor daily activities [14]. The short format of a tweet is a defined characteristic of the service, allowing informal collaboration and quick information sharing. For business, Twitter can be used to

[1] https://markets.blockchain.info/

broadcast company's latest news, posts, read comments of the customers or interact with them. A communicative feature of Twitter is the hashtag: a metatag beginning with the character #, designed to help others find a post.
Twitter is a rich source of real-time information regarding current societal trends and opinions. There are also studies that report another use of Twitter, namely as a possible predictor of market trends. Indeed, in 2010, a publication of the professor Johan Bollen showed that combining information on Wall Street with the millions of Tweets and posts, makes possible to anticipate financial performance [6]. In this work, Granger causality analysis and a Self-Organizing Fuzzy Neural Network are used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. The analysis of Tweets made by Bollen would have had 87% of chance to successfully predict prices of the stock, 3 or 4 days in advance. This study and analysis of millions of posts on Twitter represents a thermometer of emotions, on a large scale, which reflects the whole of society.
Earlier studies had found that blogs can be used to evaluate public mood, and that tweets about movies can predict box office sales. Investigating the literature related to different uses of social media, and Twitter in particular, we collected information about the use of Twitter for seeking real world emotions that could predict real financial markets trend [1]. In their paper, Rao and Srivastava investigate the complex relationship between tweet board literature (like bullishness, volume, agreement etc) with the financial market instruments (like volatility, trading volume and stock price) [2].
The Bitcoin represents an important new phenomenon in financial markets. Mai et al. [4] examine predictive relationships between social media and Bitcoin returns by considering the relative effect of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships using vector autoregressive and vector error correction models.
In their work, Garcia et al. [3] show the interdependence between social signals and price in the Bitcoin economy, namely a social feedback cycle based on word-of-mouth effect and a user-driven adoption cycle. They provide evidence that Bitcoin's growing popularity causes an increasing search volumes, which in turn result a higher social media activity about Bitcoin. More interest inspire the purchase of bitcoins by users, driving the prices up, which eventually feeds back on the search volumes.
We compared Twitter's trending topic about Bitcoin with those in other media, namely, Google Trends [8]. This is a feature of Google search engine that illustrates how frequently a fixed search term was looked for. Through this, you can compare up to five topics at one time to view relative popularity, allowing you to gain an understanding of the hottest search trends of the moment, along with those developing in popularity over time. Following this kind of approach, we evaluated how much *"bitcoin"* term, for the analyzed time interval, is looked for using Google's search engine.

## 3 Methodology

### 3.1 Sentiment Analysis

Tweets sometimes express opinions about different topics, and for this reason we decided to evaluate user's opinion about Bitcoin. We also investigated its power at predicting real-world outcomes. In order to evaluate if a user really appreciates the Bitcoin spread, we tried to predict sentiments analyzing tweets collection. In recent years, there is a wide collection of research surrounding machine learning techniques, in order to extract and identify subjective information in texts. This area is known as sentiment analysis or opinion mining [15]. Sentiment techniques are able to extract indicators of public mood directly from social media content [24].

Pang et al. argue that the research field of sentiment analysis has developed many algorithms to identify if the opinion expressed is positive or negative. In fact, algorithms to recognize sentiment are required to understand the role of emotions in informal communications [15]. Go et al. affirmed the strength of the sentiment analysis applied to the Twitter domain by using similar machine learning techniques to classifying the sentiment of tweets [19].

We chose to use automated sentiment analysis techniques to identify the sentiments of tweets in the matter of Bitcoin. Since the goal of this research is neither to develop a new sentiment analysis nor to improve an existing one, we used "SentiStrenght", a tool developed by a team of researchers in the UK that demonstrated good outcomes [11]. SentiStrength estimates the strength of positive and negative sentiments in short texts. It is based on a dictionary of sentiment words, each one associated with a weight, which is its sentiment strength. In addition, this method uses some rules for non-standard grammar.

Based on the formal evaluation of this system on a large sample of comments from MySpace.com, the accuracy of predicting positive and negative emotions was something similar to that of other systems (72.8% for negative emotions and 60.6% for positive emotions, based on a scale of 1-5). Compared to other methods, SentiStrenght showed the highest correlation with human coders [13]. The tool is able to assess each message separately and, at the end, it returns one singular value: a positive, a negative or a neutral sentiment.

### 3.2 Data Collection

Tweets are available and are easily retrieved making use of Twitter Application Programming Interface (API) [16]. Composing the hashtag #Bitcoin or @bitcoin, we are able to gather all tweets that mentioned the analyzed subject. We briefly describe the different components of our system. An overview of this architecture is shown in Figure 1. The system consists of four components:

- *Twitter Streaming API*: it provides access to Twitter data, both public and protected, on a nearly real-time basis. A persistent connection is created between our system and Twitter. As soon as tweets come in, Twitter notifies our system in real time, allowing us to store them into our database.

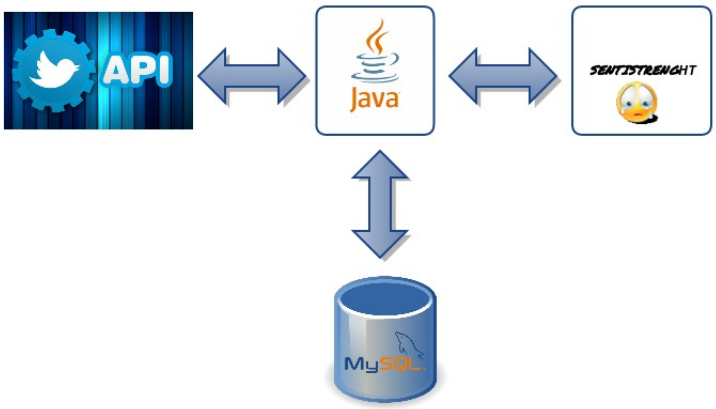


**Fig. 1.** System Architecture

- *DataStore*: our datastore consists of a back-end database engine, using MySQL as RDBMS, that repeatedly saves the incoming tweets from the Twitter Streaming API.
- *SentiStrenght tool*
- *Java Module*: this component allows us to send automated requests to Twitter Streaming API, to recover new tweets about Bitcoin, to parse the data gathered and to store them into our datastore. In a later stage, these data are sent to SentiStrenght tool in order to automatically evaluate the users' opinion.

We analyzed a collection of tweets, regarding Bitcoin, posted on Twitter between January 2015 and March 2015 (60 days). During this time 1,924,891 tweets were collected. The tweets were analyzed to determine its identifier, the date-time of the submission, its type, and its text content, which is limited to 140 characters. Comparing the timeline of tweets and the fluctuations in the Bitcoin market, we determined the specific day that provide a better correlation value. We then used SentiStrenght to evaluate comments extracted from Twitter. Given as input all tweets, the system assigned a score for each comment:

- 1 if the comment is positive
- -1 if the comment is negative
- 0 if the comment is neutral

## 4 Results

In order to decide the correct strategy of analysis for studying the relationship among Bitcoin's price and others meaningful parameters, the available related literature has been examined in depth. Most of articles [6] [1] [2] reports analysis about the existent relationship between the volume of tweets and the market evolution. In general, Bollen et al. demonstrated that tweets can predict the market trend 3-4 days in advance, with a good chance of success. We analyzed the Bitcoin price's behavior comparing its variations with the number of tweets, with the number of tweets with positive mood, and with Google Trends results. The computation of cross-correlation yielded interesting results.
Our result seems to confirm that volumes of exchanged tweets may predict the fluctuations of Bitcoin's price. Furthermore, the comparison between tweets with a positive mood and trend of Bitcoin's price seems to prove this behavior. The

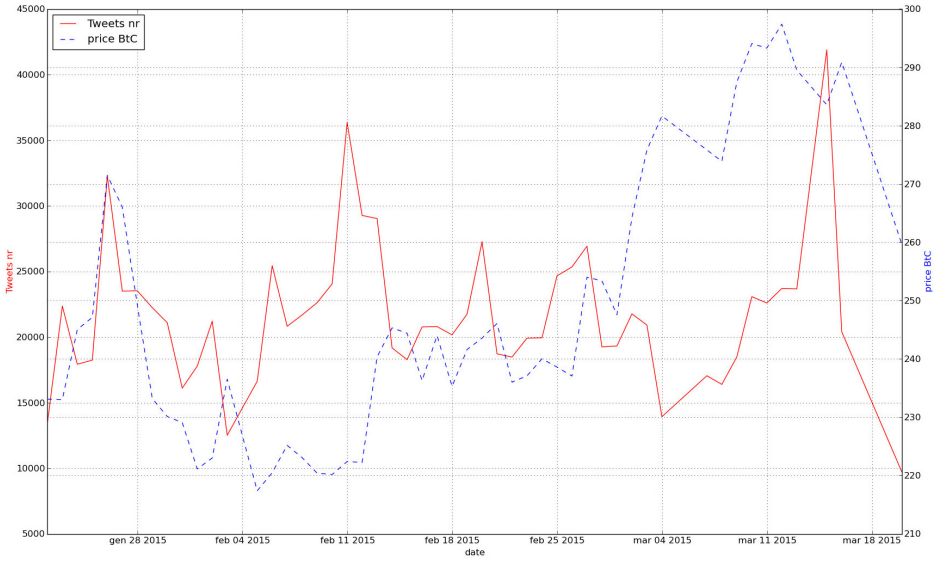


**Fig. 2.** Similarity between Bitcoin's price and number of Tweets

examined literature shows different ways to highlight the existent relationship between big volumes of exchanged tweets and meaningful variations in the Bitcoin's price. Some papers show studies using regression methodology [4] or causality analysis [6]. Rao et al.[2] and Mittal et al.[5] showed how goods and stocks markets may be influenced by a big exchanged of tweet's volume. Inspired by these works, we tried to demonstrate how chatter of tweets might predict the price's variations of Bitcoin. Figure 2 illustrates the curve trend of Bitcoin prices, expressed in dollars, and Twitter volume. We calculated the cross-correlation and,

analyzing the results, we found that, in minimal degree, tweets volume is related to price with a maximum cross correlation value of 0.15 at a lag of 1 day (this is not very significant). Nevertheless, if we observe Figure 2, we can notice how, also at a glance, there are peaks in tweets trend that precede peaks in price, suggesting a relationship between the two time series. A patent peak of tweets on 11 February, is followed by a growth of Bitcoin's price. The same circumstance is visible in the following days: January 23, February 3, February 25 and so on. We also analyzed tweets with positive mood and we noticed a two-fold increase in cross-correlation value. Figure 3 shows this result and it's well rendered that positive tweets can predict the fluctuations of the Bitcoin's price. It is proven by a maximum cross correlation value of -0.35 with a positive delay of almost 4 days. We can confirm that positive mood could predict the Bitcoin's price almost 3-4 days in advance. All patent peaks in the positive tweets plot precede a significant change in the Bitcoin's price after some days.

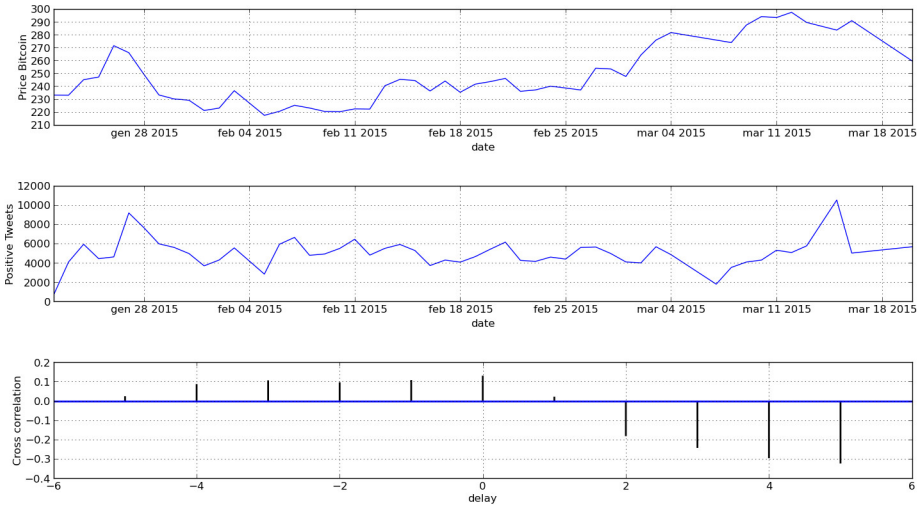


**Fig. 3.** Cross-correlation between positive Tweets and Bitcoin's price

The cross-correlation result between Google Trends data and Bitcoin's price also looks significant. The cross-correlation value increase up to a value of 0.64, that is quite substantial. This result is shown also by a little significant relationship that exists between positive tweets and Google Trends data. Figure 4 shows how Google Trends proceeds in the same direction of Bitcoin's price and highlighting a striking similarity between them. Table 1 summarizes the cross-correlation results, obtained comparing the spread among Bitcoin price and different volumes of data.

**Table 1.** *Cross-correlation results*

| Compared Systems | Cross-correlation value | delay |
|---|---|---|
| Bitcoin price-Tweets volume | 0.15 | 1 |
| Bitcoin price-Positive tweets | -0.35 | 3-4 |
| Bitcoin price-Google Trends data | 0.64 | 0 |

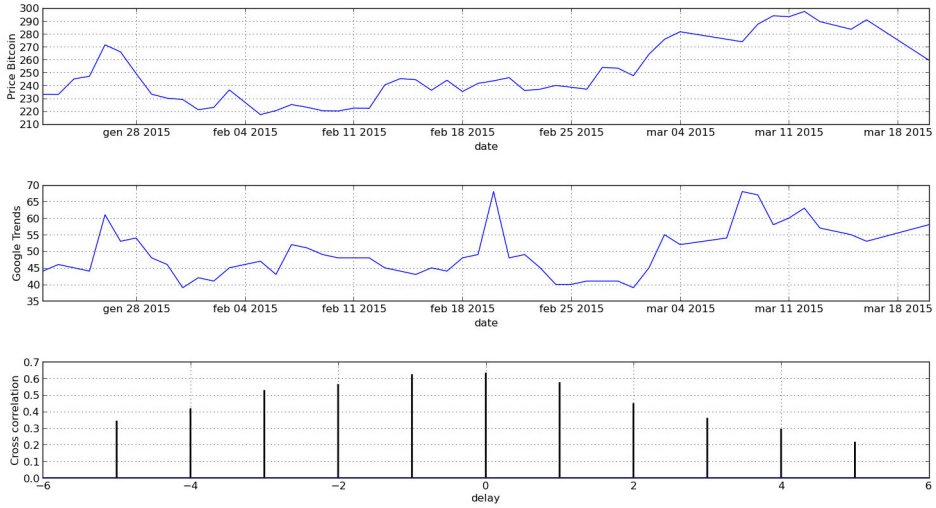


**Fig. 4.** Cross correlation between Google Trends and Bitcoin's price, expressed in dollars

## 5 Conclusions

In this paper, we studied whether social media activity or information extracted by web search media could be helpful and used by investment professionals in Bitcoins. Since the use of Bitcoins is increasingly widespread, we decided to analyze the market, in order to predict the evolution of its price.
To this purpose, we presented an analysis of a corpus of tweets about Bitcoin, considering a total amount of 1,924,891 tweets. The corpus covers a period of 60 days between January 2015 and March 2015. We applied automated Sentiment Analysis on these tweets in order to evaluate whether public sentiment could be used to predict Bitcoin's market. We also used Google Trends media to analyze Bitcoin's popularity under the perspective of Web search. In this preliminary study, we examined the Bitcoin price's behavior comparing its variations with these of tweets volume, tweets with positive mood volume and Google Trends data. From results of a cross correlation analysis between these time series, we can affirm that positive tweets may contribute to predict the movement of Bitcoin's price in a few days. Google Trends could be seen as a kind of predictor, because of its high cross correlation value with a zero lag. Our results confirm those found in the previous works, based on a different corpus of tweets and

referred to a different Bitcoin market trend.
While the current data is only 60 days already looks promising, a consecutive analysis of more than 6 months might provide a better result quality. In further studies, we also plan to take into account the number of retweets and favorites for the tweet's corpus analyzed. Along these lines, we could check whether results stay unchanged with the addition of this variable.

**Acknowledgments.** This research is supported by Regione Autonoma della Sardegna (RAS), Regional Law No. 7-2007, project CRP-17938 LEAN 2.0.

## References

1. Kaminski, Jermain, and Peter Gloor. "Nowcasting the Bitcoin Market with Twitter Signals." arXiv preprint arXiv:1406.7577 (2014).
2. Rao, Tushar, and Saket Srivastava. "Analyzing stock market movements using twitter sentiment analysis." Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society, 2012.
3. Garcia D, Tessone CJ, Mavrodiev P, Perony N. 2014 The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. J. R. Soc. Interface 11: 20140623. http://dx.doi.org/10.1098/rsif.2014.0623
4. Mai, Feng and Bai, Qing and Shan, Zhe and Wang, Xin (Shane) and Chiang, Roger H.L., From Bitcoin to Big Coin: The Impacts of Social Media on Bitcoin Performance (January 6, 2015). Available at SSRN: http://ssrn.com/abstract=2545957
5. A. Mittal, A. Goel, "Stock Prediction Using Twitter Sentiment Analysis" in proceeding of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2013
6. Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.
7. K. Dessai and M. Kamat, Application of social media for tracking knowledge in agile software projects, Available at SSRN 2018845, 2012.
8. Google Trends. `http://www.google.it/trends/` . Accessed November 6,2014.
9. Thelwall, M., Buckley, K., Paltoglou, G. : Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 62(2), 406-418 (2011).
10. Twitter [online] `https://twitter.com/` .Accessed November 6,2014.
11. SentiStrenght. [Online] `http://sentistrength.wlv.ac.uk/` .Accessed November 6,2014.
12. Alexa. Top sites: The top 500 sites on the Web. Accessed September 30, 2014, from `http://www.alexa.com/topsites/global`
13. Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., Kappas, A. : Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558 (2010).
14. Java, A., Song, X., Finin, T., Tseng, B. Why we twitter: Understanding microblogging usage and communities. Proceedings of the Ninth WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (pp. 56–65). New York: ACM Press (2007).

15. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 1(1–2), 1–135. (2008).
16. Twitter API. `https://dev.twitter.com/rest/tools/console` . Accessed November 25, 2014.
17. SOCIAL MEDIA AND MARKETS: The New Frontier
18. S. Ye and F. Wu. Estimating the size of online social networks. In Proc. of the IEEE 2nd Intl. Conf. on Social Computing (SocialCom), pages 169–176, Aug. 2010
19. A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009.
20. Grinberg, Reuben, Bitcoin: An Innovative Alternative Digital Currency, Hastings Science & Technology Law Journal, Vol. 4, p.160, Dec 2011
21. Dorit. Ron and Adi Shamir. Quantitative analysis of the full bitcoin transaction graph.
22. Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system.
23. F. Reid and M. Harrigan. An analysis of anonymity in the bitcoin system.In Proc. of the Conference on Social Computing (socialcom), 2011.Reid F.Harrigan M.
24. Pak, A and Paroubek, P. (2010) Twitter as a Corpus for Sentiment Analysis and Opinion Mining. (European Language Resources Association (ELRA), Valletta, Malta).

# Employing Relation Between Reading and Writing Skills on Age Based Categorization of Short Estonian Texts

Avar Pentel

Tallinn University, Institute of Informatics, Estonia

pentel@tlu.ee

**Abstract.** In this paper, we present results of our study on age-based categorization of short texts as 85 words per author. We introduce a novel set of features that will reliably work with short texts, and is easy to extract from the text itself without any outside databases. These features were formerly known as variables in readability formulas. We tested datasets presented two age groups - children and teens up to age 15 and adults 20 years and older. Besides readability features, we also tested widely used n-gram features. Models trained on readability features performed better or as well as models trained on n-gram features. Model generated by Support Vector Machine with readability features yield to f-score 0.953.

**Keywords:** age detection, readability features, n-grams, logistic regression, support vector machines, bayesian.

## 1 Introduction

With a wide spread of social media, growing problem is related to false identities. Younger people might pretend adults to access adult sites, and older people might pretend youngsters to communicate with youngsters. As we can imagine, this might lead to serious threats, as for pedophilia or other criminal activities. Thus, automatic age detection has serious practical application in social media.

While many works are published on text authorship profiling, social media poses two problems that are not solved this far.

The first problem is related to the amount of the text needed to make predictions. Usually a large training data sets and long texts per author are used [1,2] to make such classification models, but in social media, we can only relay on short texts.

The second problem is related to the cost of feature extraction. Most of the recent studies [3-6] on age detection using word and character n-gram based features and additional databases or systems, as part of speech tagging, etc., to assess the roles of the words in a sentence. With millions of users, these techniques are too costly to be applicable. Ideally, a system could use mostly client side resources.

In this paper, we propose a novel set of features for author's age based profiling that solves both previously mentioned problems. We call these new features the read-

ability features. These features can be easily extracted using client side JavaScript and they make at least as best classifiers as widely used n-gram based features.

We suppose that authors reading skills and writing skills are correlated, and by analyzing author's text readability, we can conclude about his/her education level, which at least to the particular age is correlated with the actual age of an author. Therefore, we can employ old readability formulas that were developed already before computerized era. Automated Readability Index [8], Gunning Fog index [9], SMOG [10], Flesch-Kincaid [11], and other readability formulas assess how much education is needed to understand particular texts. If we take a closer look at the first pair of these formulas (1,2), we can see, that they are using very simple variables, which can be easily extracted from text.

$$ARI = 4.71 \times \left(\frac{characters}{words}\right) + 0.5 \times \left(\frac{words}{sentences}\right) - 21.43 \quad (1)$$

$$GFI = 0.4 \times \left[\left(\frac{words}{sentences}\right) + 100 \times \left(\frac{complexwords}{words}\right)\right] \quad (2)$$

As readability indexes are developed for texts with about 100 words, these are good candidates for our task.

## 2 Methodology

We collected short texts, average 85 words long, from different social media sources like Facebook, Blog comments, and Internet forums. All authors were identified, and they used in their texts Estonian language. We chose balanced and stratified dataset with 400 instances and with different age groups: 7-15 and 20-48.

We used three types of features in our training datasets: readability features, character n-grams and word n-grams.

Readability features are quantitative data about texts, as for instance an average number of characters in a word, syllables in word, etc. All together 14 different features were extracted from each text as shown in Table 1.

**Table 1.** Readability features

| feature | explanation | calculation | feature | explanation | calculation |
|---|---|---|---|---|---|
| CPW | average number of characters per word | $=\frac{Characters}{Words}$ | S1TW | words with 1 syllable to all words ratio | $=\frac{1SylWords}{Words}$ |
| WPS | average number of words per sentence | $=\frac{Words}{Sentences}$ | SnTW | words with n (2-8+) syllable to all words ratio | $=\frac{nSylWords}{Words}$ |
| CPS | average number of commas per sentence | $=\frac{Commas}{Sentences}$ | CWPS | average number of complex words in sentence | $=\frac{ComplexWords}{Sentences}$ |
| SPW | average number of syllables per word | $=\frac{SyllablesInText}{Words}$ | CWTW | complex words to all words ratio | $=\frac{ComplexWords}{Words}$ |

Complex word in our feature set, is a loan from Gunning Fog Index [9], where it means words with 3 or more syllables. As in the Estonian language average number of syllables per word is higher, we raised the number of syllables accordingly. We also created a new and very simple syllable counter for Estonian language.

Another type of features we used, are character n-grams. We extracted all occurred character bigrams and trigrams and using $X^2$ attribute evaluation, we selected 119 character bigrams and 576 character trigrams.

Similarly, we extracted all occurred word unigrams, bigrams and trigrams and using $X^2$ attribute evaluation, we selected as features 100 word unigrams, 30 word bigrams and 6 word trigrams.

We made four different datasets: with readability features, with character n-grams, with word n-grams, and with all features combined. The models were generated using Support Vector Machine, Logistic Regression and Naïve Bayes algorithm. Motivation of using these algorithms comes from the literature [12]. Java implementations of listed algorithms that are available in the Weka [13] library were used. 10-fold cross validation was used for evaluation.

## 3 Results

As shown in Table 2, readability features trained a better classifier with Support Vector Machines and Logistic Regression, yielding to f-scores 0.953, and 0.95 accordingly. Naïve Bayes performed better with n-gram features. Combined feature sets did not improve the models.

**Table 2.** Results of models trained with different feature types

| Classifier | F-Scores | | | |
|---|---|---|---|---|
| | ***Readability*** | ***Char n-grams*** | ***Word n-grams*** | ***All combined*** |
| SVM standardized | 0.953 | 0.952 | 0.850 | 0.950 |
| Logistic Regression | 0.950 | 0.929 | 0.775 | 0.920 |
| Naïve Bayes | 0.811 | 0.946 | 0.901 | 0.882 |

Most distinctive features, among readability features were average number of words in a sentence and average number of characters in a word.

Using logistic regression model with readability features, we created a prototype client side age detection application [14].

## 4 Conclusion

Employing relations between reading and writing skills, and using features from old readability formulas proved to be an effective way to predict author age class. Readability features are in many ways favorable. First, they are easy to extract, they are self sufficient, and can be computed without any extra help. Syllable counting is

problematic for some languages, but maybe it can be omitted, as syllable count is also not used in all readability indexes.

Secondly, when dealing with short texts, content-based features, as n-grams tend to be very context dependent, the topic can cause a rise of frequency of some words that can be associated to a particular age group. It seems, that how we write depends less on the context than what we write.

However, we have to address limitations of the current study. First, it is obvious, that we cannot use readability features to categorize older age groups. For most of the people, reading and writing skills will not improve continuously during the whole life.

Secondly, it is possible that good age based categorization results are caused by some specific property of Estonian language. For example, Estonian language has many agglutinative inflectional suffixes, and therefore grammatical richness yield directly to more syllables and longer words. Therefore, we look forward to test how readability features work with other agglutinative and inflectional languages.

## References


1. Burrows, J. All the way through: testing for authorship in different frequency strata. Literary and Linguistic Computing. 22, 1, pp. 27–47. Oxford University Press (2007)
2. Sanderson, C., and Guenter, S. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: an investigation. EMNLP'06. Association for Computational Linguistics. pp. 482–491. Stroudsburg, PA, USA (2006)
3. Peersman, C., Daelemans, W., Vaerenbergh, L. Predicting age and gender in online social networks, SMUC '11 Proceedings of the 3rd international workshop on Search and mining user-generated contents, pp. 37-44. (2011)
4. Argamon, S., et al. Automatically profiling the author of an anonymous text. Communicatin of the ACM 52(2) pp. 119–123 (2009)
5. Nguyen, D., Rose, C.P. Age prediction from text using linear regression. LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp 115-123 (2011)
6. Weren, E.R.D. et al. Using simple sontent features for the author profiling task. Notebook for PAN at CLEF, (2013)
7. Marquart, J. et al. Age and gender identification in social media. CEUR Workshop Proceedings, vol 1180 (2014)
8. Senter, R.J., Smith, E.A. Automated Readability Index. Technical report, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio (1967)
9. Gunning, R. The Technique of Clear Writing. New York: McGraw–Hill (1952)
10. McLaughlin, G. Harry. SMOG Grading - a New Readability Formula. Journal of Reading 12 (8): 639–646 (1969)
11. Flesch, R. A new readability yardstick. Journal of Applied Psychology 32: pp. 221–233. (1948)
12. Mihaescu, M. C. Applied Intelligent Data Analysis: Algorithms for Information Retrieval and Educational Data Mining, pp. 64-111. Zip publishing, Columbus, Ohio (2013)
13. Hall, M. et al. The WEKA data mining software: an update. SIGKDD Explorations, vol 11, 1 (2009)
14. Pentel, A. Age Detector. http://www.tlu.ee/~pentel/age_detector/ (2015)

# Modeling Short-Term Preferences in Time-Aware Recommender Systems

Pierpaolo Basile, Annalina Caputo, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro

Department of Computer Science, University of Bari Aldo Moro,
Via E. Orabona 4, 70125 Bari, Italy
{name.surname}@uniba.it
http://www.di.uniba.it

**Abstract.** Recommender Systems suggest items that are likely to be the most interesting for users, based on the feedback, i.e. ratings, they provided on items already experienced in the past. Time-aware Recommender Systems (TARS) focus on temporal context of ratings in order to track the evolution of user preferences and to adapt suggestions accordingly. In fact, some people's interests tend to persist for a long time, while others change more quickly, because they might be related to volatile information needs. In this paper, we focus on the problem of building an effective profile for short-term preferences. A simple approach is to learn the short-term model from the most recent ratings, discarding older data. It is based on the assumption that the more recent the data is, the more it contributes to find items the user will shortly be interested in. We propose an improvement of this classical model, which tracks the evolution of user interests by exploiting the content of the items, besides time information on ratings. When a new item-rating pair comes, the replacement of an older one is performed by taking into account both a decay function for user interests and content similarity between items, computed by distributional semantics models. Experimental results confirm the effectiveness of the proposed approach.

**Keywords:** Time-aware Recommender Systems, Content-based Filtering, Short-Term Preferences, Distributional Semantic Models

## 1 Introduction

Recommender systems adopts information filtering algorithms to suggest items or information that might be interesting to users. In general, these systems analyze the past behavior of a user, build a model or profile of her interests, and exploit that profile to find potentially interesting items. In collaborative approaches, the user profile is usually the vector of ratings assigned to *all* items that they have accessed, viewed, or purchased [12]. Content-based approaches rely on item and user descriptions (content) to build item representations and user profiles that suggest items similar to those a target user already rated (and liked) in the past [17].

One limitation of these traditional approaches is that the temporal context of ratings is not taken in account, but actually user preferences are likely to change over time: long term interests stay stable for a long time, short term preferences tend to vary with higher frequency. There are some domains, such as news recommendation, in which retaining this temporal distinction among user preferences has an impact on the accuracy of suggestions [7, 4]. Indeed, the issue of including time information into user modeling and recommendation approaches has been investigated early in literature [3, 22], but the topic recently received renewed attention, due to the significant improvements of recommendation accuracy obtained by the time-aware algorithm adopted by the winning team of Netflix Prize competition [15].

Time-aware Recommender Systems (TARS) fall in the more general category of contex-aware ones, that exploit the context in which users express their preferences (such as: location, *time*, weather, emotional state) in order to adapt rating prediction depending on the situation in which they are experiencing an item. TARS focus on temporal context of ratings to adapt the recommendation list accordingly. Regarding the usage of time information, the literature distinguishes two classes of methods [6]:

- *time-aware* approaches, that adapt rating predictions on the target recommendation time;
- time-adaptive approaches, which do not differentiate rating predictions according to the target time, but rather adjust some parameters or data dynamically.

We focus on time-adaptive approaches, specifically on those adopting some heuristics to penalize older preferences that are presumed to be less valid at recommendation time. These methods could be considered as a particular case of time decay heuristics, but they do not target a specific recommendation time (morning, week-end, etc.). Our investigation specifically concerns approaches that adopt some time-based strategy to learn separate models for short-term preferences and for interests that persists for a long time.

In particular, in this paper, we face with the problem of building an effective short-term preference model able to predict items that will shortly be consumed by the user. To address this issue, a simple and popular approach is the use of sliding windows, that learns the model by including in the training set only the most recent ratings, while older data are discarded or weighted so that they contribute to the model in a limited way. The literature reports controversial results about the adoption of time weights for ratings provided at different times. For example, in [10] the authors show that recommendation accuracy is improved by an exponential decay function, while an opposite conclusion is drawn by performing rating prediction on the Netflix Prize dataset [14].

We argue that recent ratings of users reflect their short-term preferences more than old ratings, but this is not true for *all* old ratings, but only for *those provided on items which are different from the recently rated ones.* In other words, the hypothesis is that older ratings do not correspond definitely to older

preferences, as assumed in the classical sliding window approach, but content of the items should be taken into account as well, in order to discard old interests which differ from new ones.

The research question we want to investigate in this work is: "Is the gradual decay of influence of ratings, combined with content similarity between items, useful for modeling short-term preferences?"

We propose a time-aware distributional content-based recommender system which suggests items the user will shortly be interested in. It exploits both an exponential decay function and semantic similarity between item descriptions for regulating the item participation in building of the user profile. Items are described in a WordSpace through the geometrical metaphora of meanings. Related words are represented as near points (i.e. vectors), while the semantics of item descriptions (i.e. text fragments) is computed by summing the vectors associated with their words.

The paper is organized as follows: in Section 2 we discuss some relevant literature and compare existing approaches to the proposed one, described in the following section. Section 4 analyzes the results of the experiments performed to validate our proposal, while conclusions are drawn in the last section.

## 2 Related Work

Following the characterization given by [6], we would place our method in the class of *time-adaptive heuristic-based approaches*. Rather than modeling the notion of time as a context with respect to the items may be relevant or not, we consider time as a continuos context attribute with respect to items could be either fresh or not. Simply stated, our time-context definition aims at modelling the recency of user's preferences. This point of view has been inspired by early works by Ding et al. [10, 11], that aimed at adapting collaborative filtering algorithms in order to capture preference drift.

In [10], the authors propose a novel item-based algorithm (that identifies the similarity between two items by comparing users' ratings on them), enhanced by a time-decay function in a way that the items rated recently contribute more to the prediction of the recommendations. The underlying assumption is the same as in the sliding window approach: latest ratings reveal latest interests. The proposed approach showed the actual improvement of precision obtained by using an exponential decay function to weight ratings. In the later work [11], a recency-based approach is proposed, in which each rating of the target user is assigned a weight that is computed according to its deviation from her most recent ratings on similar items. We ground our approach on similar basis, but a significant difference is that we adopt a decay function to weigh content similarity among items recently rated and older ones, in order to select those to be included in the training set for our model. Time-adaptive heuristics have been used also by Cao [8] et al. and Lathia et al.[16]. The former approach introduces four types of user interest patterns and proposes an effective approach for detecting these patterns by exploiting user rating graphs and rating chains. The

authors show that recommendation quality improves when the derived interest patterns are taken into account. The latter formalises collaborative filtering as a time-dependent, iterative prediction problem over a dynamic, growing dataset. The authors propose adaptive temporal collaborative filtering, a method of temporally adapting the size of user kNN neighbourhoods based on the performance measured up to the current time.

Another method that adopts simple heuristics to improve collaborative filtering is proposed in [5], where the authors describe a time-biased kNN algorithm exploiting only the most recent ratings from the neighbours. It showed better performance than other kNN recommendation strategies.

Other approaches adapted factorization algorithms to face with temporal effects. Remarkable examples of these methods are described in [14] and [21]. Koren [14] suggested that a mere decay of older instances or usage of separate models for tracking the evolution of preferences cause a loss of prediction accuracy. The proposed solution is to model the temporal dynamics along the whole time period, allowing to separate volatile factors from durable ones, in order to capture the way user and product characteristics change over time. Xiong et al. [21] proposed a factorization method based on probabilistic latent factor models. In addition to the factors that are used to characterize entities, the authors introduce another set of latent features for each different time period. These additional factors represent the preference of latent features at each particular time, so that they are able to capture the evolution of preferences.

Compared to these work which handled the temporal dynamics in different ways, we focus on explicitly modeling short-term preferences and their influence on recommendation of items that will shortly be consumed.

Among approaches that propose different models for long-term and short-term preferences, Cantador et al. [7] designed a content-based news recommender system in which short-term preferences are inferred from the click history, and final ranking of items is adapted to the current context of interest. In [20], the authors propose propose a graph-based approach that introduces session nodes, associated with a user at specific time, to capture short-term linkages between items. If two items are connected by session nodes, their similarity is assumed to be contributed by short-term preferences. An algorithm for preference fusion is designed for temporal recommendation, that proved to be effective on real datasets.

Differently from previous methods, our short-term model tries to capture *semantic* similarity among items by looking at their content, rather than simply co-occurence within sessions or clicking history, with the hope that the semantic approach, combined with time information, helps to discover short-term *relatedness* among them.

## 3 Time-Aware Distributional Recommender System

Models of Distributional Semantics have drawn a lot of attention in recent years due to their capability of capturing semantics at a latent level, without the re-

quirement of learning algorithms or human-edited resources. Such models build a vector space of meanings where concepts are represented through vectors and the relatedness between meanings is expressed through a proximity function of the points they are represented by. Usually these models are built by skimming a large corpus in order to gather information about distribution of words in a text. Indeed, statistics about word co-occurrences are useful to infer paradigmatic relationships among words, i.e. relations about words that can be used interchangeably. One of the commonest use of such a model is for computing the similarity between words, since the vector components grasp the semantic of word usage in context. Then, the vector addition between words belonging to a text is an easy way to extend to a whole sentence/paragraph/document such a similarity. This work exploits the idea of adapting Distributional Semantic Model (DSM) to item descriptions as a unified framework for both representing the semantic content of items and computing the similarity between them.

### 3.1 DSM-based Recommender System

The distributional semantic-based recommender system relies on Random Indexing (RI) [13] for building up the semantic space. Given a text corpus, RI technique consists of the following two steps:

1. A *random vector* is assigned to a term in the corpus vocabulary. This vector is highly dimensional, with very few elements -called *seed*- that take values in $\{-1, 1\}$. The dimension of the reduced space corresponds to the random vector dimensionality;
2. The *semantic vector* representation for the term is built by analyzing the whole corpus and summing the random vectors of co-occurring terms in a given text window.

Mathematically, the sum over random vectors corresponds to multiply the original co-occurrence matrix by a projection operator, which preserves the distance proportion between points. The resulting space, called *WordSpace* has two strenght points: 1) the reduced dimension enables a quicker computation of similarities, and 2) likewise Latent Semantic Analysis, shrinking the number of components to a smaller set of contexts makes high order relationships more prominent.

The item space is built upon the previously computed *WordSpace*: the item representation comes from the sum of semantic vectors associated to the item textual content. The user profile, in turn, is built on the basis of the item vector representations she liked or disliked before. In particular, the model keeps trace of both positively and negatively rated items and builds two different profile vectors, $u^+$ and $u^-$, as the sum of positive and negative items, respectively. In order to get a single profile vector on which basis the model will compute the recommendations, we exploit the orthogonal projection operator, which has been successfully employed in both retrieval and recommendation scenarios [2, 18]. The idea behind the use of orthogonalization is that if two vectors are

orthogonal with respect to each other, they do not share components, which translates into "they have unrelated concepts". Hence, if we want to express the user profile through a vector that reflects the positively rated items while discarding for negative ones, logically we should represent the user vector $u$ as: $u_1^+ \vee u_2^+ \vee \ldots \vee u_n^+ \wedge NOT(u_1^-) \wedge NOT(u_2^-) \wedge \ldots \wedge NOT(u_m^-)$. However, the logical $aNOTb$ translates into a vector space endowed with a scalar product as the projection of $a$ onto the orthogonal space $\langle b \rangle^\perp \equiv \{v \in V : \forall b \in \langle b \rangle, v \cdot b = 0\}$, where $\langle b \rangle$ is the subspace $\{\lambda b : \lambda \in \mathbb{R}\}$. Thus, computing $u$ corresponds to summing all items in $u^+$ and then projecting this vector onto the orthogonal space generated by the vectors in $u^-$. However, following the De Morgan rules, the computation of $u$ can be semplified in $u = u_1^+ \vee u_2^+ \vee \ldots \vee u_n^+ \wedge NOT(u_1^- \vee u_2^- \vee \ldots \vee u_m^-)$, which corresponds to the orthogonalization of two vector (the sum of positive and negative items) performed through the Gram-Schmidt method. Then, the model of recommendation consists in exploring the set of non rated items, in order to assess their similarities with respect to the computed user profile vector. Such a similarity is computed as the cosine similarity, then the ranked list of recommended items can be presented to the user.

### 3.2 Time-adaptive algorithm

A time-adaptive user profile should be able to grasp changes in the user's behaviour in order to reflect the latest user tastes. We tackle this problem by profiling the user preferences with respect to a sliding window of time: i.e. the items which contribute to building the profile are those occurred shortly before the recommendation. This kind of short-term model was initially proposed by Billsus and Pazzani [3]. However, the user may occasionally manifest a burst of interest towards new items, then by giving more prominency only to the latest voted items can result in suggestions that diverge from the real user's preferences. The time-adaptive algorithm we propose aims to reflect the recency of user's interests in the recommendation process without completely neglecting the role of items that belong to the remote history of user. Indeed, the set of items (*profile set*) which contribute to building the user's profile is collected by taking into account two factors:

**Time:** Recent items contribute more to the profile;
**Similarity:** The profile set tends to preserve the items whose content is similar to the newly added one.

Let $I = i_1, ..., i_k$ be such a profile set, every time the user rates a new item $i_{new}$, this is automatically added to $I$, while an older item is removed. The element to be discarded is selected as follows:

$$i_{old} = argmin\{i \in I, sim(i, i_{new}) \cdot e^{-\lambda \cdot (t_{new} - t_i)}\} \quad (1)$$

where $t$ identifies the rating time, for both the new item and that under observation, $sim$ is the similarity between two items computed in the item space described in the previous subsection, and $\lambda$ is a decay factor. Equation 1 aims

to eliminate the item most dissimilar from the newly introduced one. However, in doing so we try to keep coherent the user profile by weighting this factor with the exponential function, whose role is to smooth similarity through time. Then, two possibilities may occur:

1. The new item completely diverges from the user history. In this case all items in $I$ will take on a small similarity and the exponential function will contribute more to equation 1;
2. The new item has some degree of similarity with some items in $I$. In this case the similarities will be reduced by the exponential function which serves to mediate the contribute of those similar items rated a long time ago.

Among these, the first scenario is the most interesting, since it reflects a potential new trend in the user's preference. Under this condition, the effect of equation 1 on a new incoming item would be that of consolidating this new trend in the user profile, if a newly added item is similar to that latest one, or quickly discard the "exception", thank to the contribute of the exponential function. The $\lambda$ parameter plays in this context an important role, since it regulates how fast the exponential function has to reduce its rate [10].

## 4 Evaluation

The goal of the evaluation is to assess the capability of the proposed time-adaptive algorithm to reflect the recency of user interests without loosing information about consolidated long-lasting preferences. Then, we compare our proposed algorithm (**sim**) with respect to a simple sliding window strategy (**fifo**). In **fifo**, the window of items is kept constant: as a new item is added, the older one is removed, thus following a first-in first-out strategy. This approach discards items on the basis of a mere time factor, and no information about user preferences in long-time is preserved.

### 4.1 Dataset and system setup

The evaluation is performed on the same dataset proposed by Adomavicius et al. [1] This dataset was originally designed for context-aware recommendation. However, since it contains rating timestamps, it suits our case. This dataset comprises 1,757 ratings from 117 users about 226 movies. However, after the removal of ratings without timestamps, we obtained a total of 1,492 ratings from 51 users on about 218 movies. The *WordSpace* is built by collecting co-occurrences information from a dummy corpus consisting in:

– BNC, a collection of documents from the British National Corpus (BNC)[1], containing 100 million words;

[1] http://www.natcorp.ox.ac.uk/

– CMU Movie Summary Corpus[2], a dataset of 42,306 movie plot summaries extracted from Wikipedia;
– PLOT, a collection of plots of movies in the Adomavicious corpus, extracted from Wikipedia.

The objective behind the use of such a corpus is to provide as wider as possible coverage of all word usages in a language. The *WordSpace* represents the top $150,000$ most frequent keywors, the vector dimension is set up to 400, the number of seeds is of 10 elements, while the window size for computing co-occurrences is of 5 words. The recommender system is implemented in Java, and relies on Lucene API[3] for building both the *WordSpace* and the recommendation model. The factor $\lambda$ in equation 1 is set to 0.01.

### 4.2 Evaluation protocol

We evaluate the proposed model on a top-$N$ recommendation task. The evaluation is performed as a time-dependent cross-validation, based on increasing time window [6]. This means that the dataset is split on the basis of temporal order of rated items. For each user, we order ratings on the temporal line, then we choose the first $k_1$ elements as training and the following $k_2$ items as test set, where $N$ has to be chosen $\leq k_2$. Then, this methods computes iteratively the training and test set by adding the previous test to the current training set, while the new training set is made by sliding $k_2$ items along the temporal axis. Each iteration corresponds to a single fold. Then, the evaluation metrics are computed over all the users. We compare the **sim** approach with respect to the **fifo** baseline in terms of Mean Average Precision (MAP) [9] and NDCG (Normalized Discounted Cumulative Gain) [19], two metrics that are particularly suitable for our evaluation since they take into account the order of items in the final rank. In fact, the goal of the evaluation is to assess the ability of the proposed method to suggests items which will shortly be consumed by the user. Therefore, the ranking computed by our recommendation method is compared to the *ideal* ranking defined in the test set by ratings and corresponding time information. In other words, we want to evaluate whether our method is able to rank in the top positions of the recommendation list those items in the test set having high ratings and next to the training set, along the temporal axis.

### 4.3 Analysis of the Results

We compared the two time-adaptive methods on a variable $k_1$, i.e. the training set dimension. We set $k_2 = 5$ and $N = 3$, while varing $k_1 \in \{5, \ldots, 15\}$. We decided to use this strategy in order to assess the validity of the proposed method when a wider profile set is available.

Tables 1a and 1b report the results of the evaluation with respect to MAP and NDCG metrics. Both tables show a similar trend. The best result is obtained on

[2] http://www.ark.cs.cmu.edu/personas/
[3] http://lucene.apache.org/

$k_1 = 7$ by **sim** method, which gives also better overall figures. There are only two exceptions to this trend, with $k_1 = 11$ and $k_1 = 14$, although the differences in reported values are very small. Another common trend showed by results is that, as we increase the $k_1$ dimension, the differences between **sim** and **fifo** become smaller. However, we ascribe such a trend to the fact that by increasing the dimension of the training set, the differences between **fifo** and **sim** strategies become smaller since the variability in the set decreases (i.e. an increasingly number of items from the past of the user become part of the training set).

Table 1: Evaluation results at different size of $k_1$.

(a) MAP

| $k_1$ | **fifo** | **sim** |
|---|---|---|
| 5 | 0.447 | 0.460 |
| 6 | 0.444 | 0.469 |
| 7 | 0.465 | 0.499 |
| 8 | 0.456 | 0.461 |
| 9 | 0.459 | 0.469 |
| 10 | 0.457 | 0.466 |
| 11 | 0.458 | 0.455 |
| 12 | 0.459 | 0.468 |
| 13 | 0.449 | 0.457 |
| 14 | 0.460 | 0.458 |
| 15 | 0.431 | 0.451 |

(b) NDCG.

| $k_1$ | **fifo** | **sim** |
|---|---|---|
| 5 | 0.555 | 0.569 |
| 6 | 0.552 | 0.575 |
| 7 | 0.569 | 0.597 |
| 8 | 0.559 | 0.564 |
| 9 | 0.560 | 0.575 |
| 10 | 0.566 | 0.573 |
| 11 | 0.562 | 0.557 |
| 12 | 0.558 | 0.565 |
| 13 | 0.550 | 0.557 |
| 14 | 0.567 | 0.564 |
| 15 | 0.537 | 0.555 |

## 5 Conclusion and Future Work

Generally, the behavior of a user may be determined by her long-term interests, but at any given time, she is also affected by short-term preferences or information needs. In this paper, we argue that time is not the only factor to be taken into account in order to distinguish among short-time and long-time interests. We started from the classical sliding window approach and suggested that older ratings do not correspond definitely to older preferences, but content of the items should be considered as well. We proposed an approach that models short-term preferences by adopting a content-based sliding window approach: when a new ratings comes into the system, the replacement of an older one is performed by taking into account both a decay function for user interests and content similarity between items on which ratings are provided, computed by distributional semantics models. We compared the proposed approach to the simple FIFO strategy (the new rating replaces the oldest one). Experimental results confirmed the hypothesis that the gradual decay of influence of ratings,

combined with content similarity between items, is actually useful for modeling short-term preferences, especially when a few ratings are available to train the system. As a future work, we plan to evaluate our short-term model on a wider dataset. Furthermore, we want to design a model for long-term preferences, as well as a way to integrate the two models in order to have an overall recommendation strategy.

## Acknowledgements

This work fulfils the research objectives of the PON 02_00323_2938699 CUP B86D130000300007 project "EFFEDIL - SOLUZIONI INNOVATIVE PER L'EFFICIENZA ENERGETICA IN EDILIZIA" funded by the Italian Ministry of University and Research (MIUR).

## References

1. Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating contextual information in recommender systems using a multidimensional approach. ACM Trans. Inf. Syst. 23(1), 103–145 (Jan 2005)
2. Basile, P., Caputo, A., Semeraro, G.: Negation for document re-ranking in ad-hoc retrieval. In: Amati, G., Crestani, F. (eds.) Advances in Information Retrieval Theory, Lecture Notes in Computer Science, vol. 6931, pp. 285–296. Springer Berlin Heidelberg (2011)
3. Billsus, D., Pazzani, M.J.: User modeling for adaptive news access. User Model. User-Adapt. Interact. 10(2-3), 147–180 (2000), http://dx.doi.org/10.1023/A:1026501525781
4. Billsus, D., Pazzani, M.J.: The adaptive web. chap. Adaptive News Access, pp. 550–570. Springer-Verlag, Berlin, Heidelberg (2007), http://dl.acm.org/citation.cfm?id=1768197.1768218
5. Campos, P.G., Bellogín, A., Díez, F., Chavarriaga, J.E.: Simple time-biased knn-based recommendations. In: Proceedings of the Workshop on Context-Aware Movie Recommendation. pp. 20–23. CAMRa '10, ACM, New York, NY, USA (2010)
6. Campos, P., Dez, F., Cantador, I.: Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. User Modeling and User-Adapted Interaction 24(1-2), 67–119 (2014), http://dx.doi.org/10.1007/s11257-012-9136-x
7. Cantador, I., Bellogín, A., Castells, P.: Ontology-based personalised and context-aware recommendations of news items. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01. pp. 562–565. WI-IAT '08, IEEE Computer Society, Washington, DC, USA (2008), http://dx.doi.org/10.1109/WIIAT.2008.204
8. Cao, H., Chen, E., Yang, J., Xiong, H.: Enhancing recommender systems under volatile userinterest drifts. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. pp. 1257–1266. CIKM '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1645953.1646112
9. Croft, B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company, USA, 1st edn. (2009)

10. Ding, Y., Li, X.: Time weight collaborative filtering. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management. pp. 485–492. CIKM '05, ACM, New York, NY, USA (2005), http://doi.acm.org/10.1145/1099554.1099689
11. Ding, Y., Li, X., Orlowska, M.E.: Recency-based collaborative filtering. In: Proceedings of the 17th Australasian Database Conference - Volume 49. pp. 99–107. ADC '06, Australian Computer Society, Inc., Darlinghurst, Australia, Australia (2006), http://dl.acm.org/citation.cfm?id=1151736.1151747
12. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA. pp. 230–237. ACM (1999), http://doi.acm.org/10.1145/312624.312682
13. Kanerva, P.: Sparse Distributed Memory. MIT Press (1988)
14. Koren, Y.: Collaborative filtering with temporal dynamics. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 447–456. KDD '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1557019.1557072
15. Koren, Y.: Collaborative filtering with temporal dynamics. Commun. ACM 53(4), 89–97 (Apr 2010), http://doi.acm.org/10.1145/1721654.1721677
16. Lathia, N., Hailes, S., Capra, L.: Temporal collaborative filtering with adaptive neighbourhoods. In: Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 796–797. SIGIR '09, ACM, New York, NY, USA (2009)
17. Lops, P., de Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 73–105. Springer (2011), http://dx.doi.org/10.1007/978-0-387-85820-3_3
18. Musto, C., Semeraro, G., Lops, P., de Gemmis, M.: Random indexing and negative user preferences for enhancing content-based recommender systems. In: Huemer, C., Setzer, T. (eds.) E-Commerce and Web Technologies, Lecture Notes in Business Information Processing, vol. 85, pp. 270–281. Springer Berlin Heidelberg (2011)
19. Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 257–297. Springer US (2011)
20. Xiang, L., Yuan, Q., Zhao, S., Chen, L., Zhang, X., Yang, Q., Sun, J.: Temporal recommendation on graphs via long- and short-term preference fusion. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 723–732. KDD '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1835804.1835896
21. Xiong, L., Chen, X., Huang, T., Schneider, J.G., Carbonell, J.G.: Temporal collaborative filtering with bayesian probabilistic tensor factorization. In: Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA. pp. 211–222. SIAM (2010), http://dx.doi.org/10.1137/1.9781611972801.19
22. Zimdars, A., Chickering, D.M., Meek, C.: Using temporal data for making recommendations. In: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence. pp. 580–588. UAI '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001), http://dl.acm.org/citation.cfm?id=647235.720264