Frontmatter

Editors

SIGAPP FY’16 Quarterly Report

J. Hong

Selected Research Articles

On the Quality of Semantic Interest Profiles for Online Social Network Consumers

C. Besel, J. Schlötterer, and M. Granitzer

Using Mobile Messages to Improve Student Participation in Blended Courses: A Brazilian Case Study

E. Marçal, R. Andrade, R. Melo, W. Viana, and E. Junqueira

Discriminating Graph Pattern Mining from Gene Expression Data

F. Fassetti, S. Rombo, and C. Serrao

Multimodal Human Attention Detection for Reading from Facial Expression, Eye Gaze, and Mouse Dynamics

J. Li, G. Ngai, H. Leong, and S. Chan
SIGAPP FY’16 Quarterly Report

July 2016 – September 2016
Jiman Hong

Mission

To further the interests of the computing professionals engaged in the development of new computing applications and to transfer the capabilities of computing technology to new problem domains.

Officers

<table>
<thead>
<tr>
<th>Role</th>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chair</td>
<td>Jiman Hong</td>
<td>Soongsil University, South Korea</td>
</tr>
<tr>
<td>Vice Chair</td>
<td>Tei-Wei Kuo</td>
<td>National Taiwan University, Taiwan</td>
</tr>
<tr>
<td>Secretary</td>
<td>Maria Lencastre</td>
<td>University of Pernambuco, Brazil</td>
</tr>
<tr>
<td>Treasurer</td>
<td>John Kim</td>
<td>Utica College, USA</td>
</tr>
<tr>
<td>Webmaster</td>
<td>Hisham Haddad</td>
<td>Kennesaw State University, USA</td>
</tr>
<tr>
<td>Program Coordinator</td>
<td>Irene Frawley</td>
<td>ACM HQ, USA</td>
</tr>
</tbody>
</table>

Notice to Contributing Authors

By submitting your article for distribution in this Special Interest Group publication, you hereby grant to ACM the following non-exclusive, perpetual, worldwide rights:

- to publish in print on condition of acceptance by the editor
- to digitize and post your article in the electronic version of this publication
- to include the article in the ACM Digital Library and in any Digital Library related services
- to allow users to make a personal copy of the article for noncommercial, educational or research purposes

However, as a contributing author, you retain copyright to your article and ACM will refer requests for republication directly to you.

Next Issue

The planned release for the next issue of ACR is December 2016.
On the Quality of Semantic Interest Profiles for Online Social Network Consumers

Christoph Besel  
University of Passau  
Innstraße 41  
Passau, Germany  
christoph.besel@google mail.com

Jörg Schlötterer  
University of Passau  
Innstraße 41  
Passau, Germany  
joerg.schloetterer@uni-passau.de

Michael Granitzer  
University of Passau  
Innstraße 41  
Passau, Germany  
michael.granitzer@uni-passau.de

ABSTRACT
Social media based recommendation systems infer users’ interests and preferences from their social network activity in order to provide personalised recommendations. Typically, the user profiles are generated by analysing the users’ posts or tweets. However, there might be a significant difference between what a user produces and what she consumes. We propose an approach for inferring user interests from followees (the accounts the user follows) rather than tweets. This is done by extracting named entities from a user’s followees using the English Wikipedia as knowledge base and regarding them as interests. Afterwards, a spreading activation algorithm is performed on a Wikipedia category taxonomy to aggregate the various interests to a more abstract and broader interest profile. We evaluate the coverage of followee lists in terms of named entities and show that they provide sufficient input to infer comprehensive semantic interest profiles. Further, we compare the profiles created with the followee-based approach against tweet-based profiles. With over 7 out of 10 items being relevant to the users in our evaluation, we show that the followee-based approach can compete with the state of the art and performs even better in predicting the users’ interests than their human friends do.

CCS Concepts
•Information systems → Personalization; •Human-centered computing → Social networks;

Keywords
Personalization, Twitter User Profile

1. INTRODUCTION
We have seen a rapid increase in the amount of published information and data since the rise of the Internet. Obviously, it is not possible for humans to process all the information available, a problem known as “information overload” [5]. At the same time more and more people reveal their interests explicitly in and implicitly by using social networks. The goal of social media based recommendation systems is to infer users’ interests and preferences from their social network activity and use the thereby generated interest profiles for making personalized content recommendations. Using social information for recommendation systems is also connected to the hope of solving the cold start problem which in particular correlation based approaches suffer from, especially for smaller web pages. The cold start problem concerns the issue that a system does not know anything about new users and needs an initial phase to gather information about them. Most of the related work infers the interest profiles from a user’s posts or tweets. However, there might be a significant difference between what a user produces and what she consumes. Moreover the passive use of social network sites is on the rise. Now four in ten users browse Facebook only passively, without posting anything [6]. For those users, profile construction based on a user’s postings fails, since there is simply no input from which the profile could be created. We address this problem by inferring semantic interest profiles from the twitter followees (the accounts, the user follows) rather than her tweets. It is to note, that while we focus on Twitter and followees, the approach could be adapted to other online social networks as well, by accounting for the corresponding features, e.g. likes on Facebook.

The rationale for the followee-based approach is that many famous people maintain a Twitter account and a lot of Twitter users follow these accounts. For those accounts, the likelihood that a Wikipedia article about this person exists is very high. Moreover, Wikipedia articles are typically linked to higher level categories (e.g., the article about the football player “Thomas Müller” is linked to the category “German footballers”). Making use of those categories, following an account that can be linked to a Wikipedia article can be seen as implicit expression of interests (e.g., following the football player “Thomas Müller” reveals interest in “German footballers”). In addition, the assigned categories are organised in some kind of hierarchy in Wikipedia, thus they can be traversed in order to provide a more fine- or coarse-grained profile. This approach immediately raises the question of whether a sufficient number of followees can be linked to Wikipedia entities, which we address in the first part of the paper.

Specifically, the contribution of this paper is the following:
• We evaluate the coverage of followee lists in terms of named entities in the English Wikipedia and show that the followee lists provide enough input to infer comprehensive semantic interest profiles.

• We propose a followee-based approach to create user interest profiles, which can compete with state of the art tweet-based approaches.

• We compare the similarity of followee- and tweet-based profiles and show that they are more similar on very concrete and abstract levels than in between.

The remainder of the paper is organized as follows: In the next section, we present related work in the field of social media based recommendation systems. Then we provide an overview of the approach, followed by the evaluation of named entity coverage in followee lists and the evaluation of the overall quality of the approach by a user study. In the last part, we compare and analyze the similarity of profiles generated with the proposed approach against tweet-based profiles. Finally, we conclude the paper and provide an outlook on future work.

2. RELATED WORK

Research on user profiling and personalized content recommendation has been done for many years since the beginning of the web [10]. Early approaches focused on the web [10] and search history [18] of the user. Recently, with the emergence of social networks like Twitter, research has shifted to analyze user activities on these platforms. For instance Siehndel and Kawase [17] introduced TwikiMe, a prototype for generating user profiles by extracting entities from the user’s tweets and linking them to the 23 top-level categories of the English Wikipedia. This leads to abstract interest profiles with a fixed size represented as a 23-length vector.

Abel et. al. [1] in their work compared hashtag-based, topic-based (bag-of-words) and entity-based user models generated from the user’s tweets, for news recommendation. In this approach the scoring of the extracted concepts and interests is based on a simple term frequency technique. The results of their comparative evaluation showed that the simple bag-of-words and hashtag-based approaches, which did not consider the semantics of a tweet, were clearly outperformed by the (semantic) entity-based approach (precision of 0.71 compared to 0.4 and 0.1). Based on these results Tao et. al. [19] presented TUMS, a Twitter-based User Modeling Service, that tries to infer semantic user profiles from the messages people post on Twitter. However, the focus of TUMS is to make use of semantic web technologies for providing a standardized representation of the interest profiles allowing an easy exchange between different web services. This is connected with the hope to solve the so-called ramp up or cold start problem, a downside of approaches like content based or collaborative filtering [19] [12], which usually depend on the build-up of a user history before making personalized content recommendations. In terms of the applied algorithm and the knowledge base, the approach introduced by Kapin pathi et. al. [2] is the closest to our work. They used the English Wikipedia to spot entities in tweets and leveraged the hierarchical relationships by performing a spreading activation on the Wikipedia Category Graph to infer user interests. The result, a weighted hierarchical interest profile (expressed as a so-called Hierarchical Interest Graph), was evaluated by a user study which showed an average of approximately eight out of the ten interests in the graph being relevant to a user.

Even though Siehndel and Kawase [17] suggested investigating other types of inputs for inferring user interests, most of the related work only makes use of the content posted by a user (e.g. the tweets). Some approaches tried to consider the social graph of the user at least to some extent whereas Lim and Datta [11] presented a basic approach for interest profile creation based on celebrities, a user follows. These celebrities are classified as belonging to one or more of 15 predefined interest categories. The classification is based on the celebrity’s occupation field on his or her Wikipedia page and a set of keywords associated with each interest category. While this approach is also based on followees, in contrast to our work, it ignores the category information provided by Wikipedia and provides only support for a fixed (and predefined) set of interests, similar to Siehndel and Kawase [17].

Most similar to our work, a recent approach by Faralli et al. [3] also utilizes followees and the Wikipedia Bitaxonomy. However, while there are similarities in the applied approach, Faralli et al. do not directly evaluate the semantic interest profiles. Instead, they use it to identify users as belonging to a target population or not. Further, they apply itemset mining and based on the itemsets and association rules, they provide recommendations, e.g. for topical friends or categories a user might also be interested in and evaluate those recommendations. We in contrast evaluate whether the constructed profile really describes the user.

3. APPROACH OVERVIEW

The generation of interest profiles in this paper can be seen as a four-step process which is shown in fig. 1. In the following, each step is described in more detail and a fictional user called @soccerfan will be used as an illustrating example.

Fetch user’s friends In the first step the accounts which are followed by the user (the followees) are crawled. This is done through Twitter’s RESTful Web API[4]. As the API applies strict rate limits, extensive use of caching techniques is made to reduce the number of requests sent to Twitter.

The fictional user @soccerfan might, among others, follow the accounts @Cristiano (Cristiano Ronaldo), @BSchweinsteiger (Basti Schweinsteiger), @neymarjr (Neymar Jr), @esmuellert (Thomas Müller) and @FIFAcom (FIFA.com).

Link friends to entities The objective of this step is to link the user’s followees to corresponding entities represented by Wikipedia articles. This entity linking includes handling coincidental homonymy and ambiguity (for instance there are several famous “Thomas Müllers” with their own Wikipedia page). For that

interest profile

1 Fetch user’s friends

2 Link friends to entities

3 Aggregate to interest profile

4 Output and representation

Twitter API

Wikipedia API

WiBi Taxonomy

Figure 1: Overview approach

The extracted Wikipedia article entities are assigned to Wikipedia categories. These categories are hierarchically structured (at least to some extent) and used to represent particular interests of the user. By performing a spreading activation algorithm on the Wikipedia Bitaxonomy (a taxonomy based on the Wikipedia page and category hierarchy) the single interest entities are aggregated to a more abstract and broader interest profile.

The categories of the Wikipedia page entities extracted in the previous step represent the set of initially activated nodes. Their activation is spread during several iterations to neighboring nodes connected by outgoing edges. Formally the activation $a(v)$ of a node $v$ can be written as:

$$a_t(j) \leftarrow a_{t-1}(j) + d \cdot a_{t-1}(i)$$

where $j$ is being activated by node $i$ and $0 < d < 1$ represents the decay factor. If a node is activated by more than one node the activation is accumulated in this node. Apart from that, a normalization with the number of incoming edges and a so-called Intersection Boost (see [7] for more details), boosting nodes that are intersections of different paths are applied.

In our example the entities (pages) are assigned to categories such as 2014 FIFA World Cup players or German footballers. Performing spreading activation identifies sports and footballers as two of the most suitable overall interest categories for the example user.

Output and representation As the output of step three is a graph data structure with weighted nodes, the objective of this last step is to convert this representation to a common exchange format. Therefore, the top-k interests are extracted and can be represented in an arbitrary format. Typical representations include JSON or XML and semantic web vocabularies, such as the FOAF (Friend of a Friend) or Weighted Interests Vocabulary could be used. This also allows the provision of the interest profiles to other applications and web services through standardized interfaces.

4. ENTITY COVERAGE EVALUATION

The first question we need to address is whether the followee list of a Twitter user is sufficient input for inferring his or her interest profile. This mainly depends on the number of followees which could be linked to an entity and the quality of that entity linking. We evaluated both issues on a sample dataset.

4.1 Method and sample description

We conducted experimental research by crawling the profiles of 3000 twitter accounts (with over 350 000 followees in total) chosen randomly from an updated data set based on [2, 8]. Afterwards we analyzed the number of followees that could be linked to an entity and assessed the quality of that entity linking by applying the disambiguation heuristics mentioned in the second step of section 3.

A first analysis of the sample showed that over 72% of the users in the sample are friends with more than 50 other accounts. More than half of the Twitter accounts examined had between 50 and 200 followees. The overwhelming majority (91%) used the English language version of Twitter.

4.2 Quantitative results

For analyzing the number of followees that could be linked to a corresponding Wikipedia page entity we used the MediaWiki Web API. As this API allows search on the English Wikipedia with an auto suggest feature enabled or disabled, we did the calculation for both. Table 1 and table 2 show
the results for different selections on the sample. The numbers include the shares of followees which could be linked to an entity unambiguously, the followees that could be linked to more than one page (ambiguity) and the followees that could not be linked to any entity at all.

Table 1: Quantitative results (auto suggest enabled)

<table>
<thead>
<tr>
<th>Selection</th>
<th>unambiguously</th>
<th>ambiguously</th>
<th>not at all</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>69.89</td>
<td>7.14</td>
<td>22.72</td>
</tr>
<tr>
<td>Number of followees &gt; 50</td>
<td>71.08</td>
<td>7.11</td>
<td>21.65</td>
</tr>
<tr>
<td>Number of followees &lt; 50</td>
<td>66.77</td>
<td>7.23</td>
<td>25.51</td>
</tr>
<tr>
<td>English language version</td>
<td>71.24</td>
<td>7.24</td>
<td>21.27</td>
</tr>
<tr>
<td>Other language version</td>
<td>54.84</td>
<td>6.05</td>
<td>38.87</td>
</tr>
<tr>
<td>English language version, number of followees &gt; 50</td>
<td>72.44</td>
<td>7.20</td>
<td>20.22</td>
</tr>
</tbody>
</table>

On average about 70\% of the total number of followees could be linked unambiguously to an entity by the MediaWiki API with the auto suggest feature enabled. In less than every tenth case (7.14\%) more than one disambiguation (articles of the same name) was possible. About a fifth of the followees could not be linked to any entity even with the auto suggest feature enabled. Considering only accounts using the English language version the share of followees linked unambiguously is significantly higher (71.24\%) than with other language versions (54.84\%). The same effect, though to a lesser extent, can be seen when comparing accounts that have more and less than 50 followees. The best success rate (72.44\%) is achieved by a combined selection of accounts using the English language version of Twitter with more than 50 followees.

Table 2: Quantitative results (auto suggest disabled)

<table>
<thead>
<tr>
<th>Selection</th>
<th>unambiguously</th>
<th>ambiguously</th>
<th>not at all</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>41.23</td>
<td>5.73</td>
<td>52.93</td>
</tr>
<tr>
<td>Number of followees &gt; 50</td>
<td>42.74</td>
<td>5.81</td>
<td>51.35</td>
</tr>
<tr>
<td>Number of followees &lt; 50</td>
<td>37.24</td>
<td>5.54</td>
<td>57.08</td>
</tr>
<tr>
<td>English language version</td>
<td>42.61</td>
<td>5.88</td>
<td>51.39</td>
</tr>
<tr>
<td>Other language version</td>
<td>25.84</td>
<td>4.06</td>
<td>69.99</td>
</tr>
<tr>
<td>English language version, number of followees &gt; 50</td>
<td>44.17</td>
<td>5.95</td>
<td>49.79</td>
</tr>
</tbody>
</table>

Table 3: Qualitative results (overlap coefficient)

<table>
<thead>
<tr>
<th>Selection</th>
<th>Entity Linking</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 7500</td>
<td>n = 7500</td>
</tr>
<tr>
<td>no normalization</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>0.2325</td>
<td>0.0910</td>
<td>0.2168</td>
</tr>
<tr>
<td>normalization</td>
<td>0.0609</td>
<td>0.0643</td>
</tr>
</tbody>
</table>

With auto suggest feature disabled the share of followees that could be linked to an entity is, as one could expect, lower (41.23\% compared to 69.89\%). However the trends for the different sections are very similar. For accounts with more than 50 followees that use the English language version barely half could be linked to an entity (6\% of these ambiguously).

4.3 Qualitative results

The quantitative results may not necessarily imply that the quality of the entity linking is sufficient. This depends on whether the followee was linked with the semantically correct entity. For instance “common” people that share the name with a celebrity coincidentally might be linked to a Wikipedia page. To assess the quality of the entity linking we applied some of the disambiguation heuristics mentioned in section 3.

4.3.1 Overlap coefficient

Even though the applicability for tweets might be limited due to their short length and informal character we first calculated the overlap coefficient as a simple syntactic measure for assessing the link quality. This was done by collecting the last 20 tweets for 7500 randomly chosen Twitter users that could be linked to a Wikipedia page by the MediaWiki Web API (auto suggest enabled) and the summary of the linked page (usually the very first section). Afterwards we tokenized the crawled input and converted it into a set of words, which also removed duplicates. On that basis we calculated the overlap coefficient as shown in eq. (2) (where X and Y are the two word token sets compared).

\[
\text{overlap}(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)}
\] (2)

The overlap coefficient was calculated for both, a random mapping of tweets and page summaries (the baseline) and for the linking suggested by the MediaWiki Web API. This was done before a text normalization (stop word removal and stemming) was applied as well as afterwards. With text normalization the results show (see table 3) that the mean overlap coefficient for the entity linking is twice as high as for the random baseline mapping (Cohen’s d = 0.47). Without text normalization the effect (Cohen’s d = 0.18) is clearly smaller. All value differences were highly statistically significant (p < 0.001).
4.3.2 Reverse Linking

A very easy way to get a quick estimation of the entity linking quality is to search on Twitter for accounts with the name of the Wikipedia page title (entity). By doing this reverse linking we ended up in about 80% of the cases with the account we started the entity linking from. Both Twitter and Wikipedia have optimized search indices and the article title sometimes contains additional disambiguation information (e.g. “footballer” for “Thomas Müller”). A high success rate could be seen as indicator of a good entity linking, but as the search algorithms and indices of Twitter and Wikipedia are black boxes for us this could only be a first clue.

4.3.3 Sense Prior

Sense Prior is a probabilistic approach which assumes that the most frequent word meaning dominates the others [16]. For that purpose the relative frequencies of so-called surface forms linking to an entity are calculated and the most frequent one is assumed to be the correct disambiguation. In this evaluation we used a dataset based on the internal link structure of the English Wikipedia to calculate the frequencies. Let \( l = (s, a) \) be an internal link which points to article \( a \) with the link text (surface form) \( s \) and let \( n(s) \) describe the number of occurrences of that surface form in all articles then

\[
P(a|s) = \frac{l(s, a)}{n(s)}
\]

is the probability that entity \( a \) is the correct disambiguation for surface form \( s \).

We calculated that probability for 10 000 randomly chosen followee names (the surface form in this case) of our sample (no auto suggest, English language version and more than 50 followees) and compared the entity with the highest probability to the linked entity. If the Sense Prior dataset could provide a disambiguation (the case in 78%) it corresponded with a probability of over 90% with the linked entity.

4.4 Analysis and Discussion

The results of our empirical research show that without auto suggest almost half of the followees and with auto suggest over two thirds of the accounts a user is following could be linked to Wikipedia page entities successfully. This implies that the Twitter followees of a user actually could be a sufficient and broad basis for inferring interest profiles. As ambiguity does occur only in about one out of ten cases it should have little effect.

The qualitative evaluation points towards the same direction: With an overlap coefficient twice as high as for a random baseline mapping and success rates of about 80% for the reverse linking and around 80% and 90% for the probabilistic disambiguation heuristics the entity linking quality could be seen as sufficient as well.

We conclude, that the Twitter followees of a user provide already a sufficient input, both quantitatively and qualitatively, for inferring meaningful interest profiles.

5. USER STUDY

Even though the groundwork in the last section showed that the Twitter followees are a sufficient base for inferring interest profiles, the evaluation of personalization and/or recommendation systems typically involves a user study [7]. For that purpose we implemented the approach presented in section 4 in Python and evaluated it with real users. The modular application made use of several external modules such as tweepy for accessing the Web APIs and networkx for building the taxonomy graph and performing spreading activation on it. The source code of the application can be found on the project repository [8].

Table 4: Evaluated profile types

<table>
<thead>
<tr>
<th>Profile type</th>
<th>Iterations</th>
<th>Decay</th>
<th>Disambiguation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile type 1</td>
<td>5</td>
<td>0.2</td>
<td>No</td>
</tr>
<tr>
<td>Profile type 2</td>
<td>5</td>
<td>0.2</td>
<td>Yes</td>
</tr>
<tr>
<td>Recommendations</td>
<td>3</td>
<td>0.2</td>
<td>Yes</td>
</tr>
<tr>
<td>Comparative Eval.</td>
<td>5</td>
<td>0.2</td>
<td>Yes</td>
</tr>
</tbody>
</table>

After the users had registered by providing their Twitter screenname and e-mail, they were notified by a mail providing a link to their personalized questionnaire. This questionnaire had four pages that corresponded with the four different interest profile types shown in table 4. For screenshots of the registration form and questionnaire pages please see the project repository [9].

On the first page the user was presented the top 20 interest categories (most weighted nodes) of the first profile type. The participants were asked to indicate their strength of interest for each category on a four-point Likert scale ranging from “very interesting” to “not interesting at all”.

The second page was pretty much the same presenting the top 20 interests of profile type 2 that mainly differed in whether disambiguation heuristics were applied or not. On the third page five Wikipedia articles that were assigned to the interest categories of the third interest profile (a smaller number of iterations was used to get more specific results) were shown. Again the participants were asked to indicate their strength of interest in the topics covered by these articles.

The last page showed the users ten interest categories randomly picked from the profiles of other users. As the categories did not appear in their interest profiles, no interest

8 http://www.tweepy.org/
9 https://pypi.python.org/pypi/wikipedia/
10 https://networkx.github.io/
11 The project repository includes application source code, evaluation scripts and screenshots https://bitbucket.org/beselch/interest_twitter_acmsac16

5.1 Experimental Setup

For our evaluation we generated four different profile types defined by the number of iterations, the decay factor and the application of disambiguation heuristics (see table 4).
of the users in these categories was assumed and they were asked to evaluate whether this was correct or not.

Afterwards the participants were provided with a link they were asked to send a friend of theirs. This link lead to a one-paged survey that presented the user’s friend with 20 interest categories. One half consisted of the top 10 interests of profile type 2 and the other half were the randomly picked interest categories that our approach assumed to be not interesting. Now the user’s friend was asked to evaluate the interest of his or her friend in these interest categories. Following [20] these answers were used to compare the performance of the friend and our algorithm in predicting the user’s interests. Whereas the pages one and two were obligatory the last two steps could be skipped by the participants.

5.2 Sample description

During the evaluation period from 30 June to 10 July 2015 64 Twitter users registered for the user study and 52 of them completed the survey (response rate of 81.25 %). A participant had 205 followees on average, while the median (114) was considerably lower. The used Twitter language versions were half German and half English. Barely half of the users posted fewer than 100 tweets (over 15 % nothing), which means that approaches based on the tweets would fail to generate interest profiles for that users. 46 participants submitted the optional third page and 17 people took part in the fourth step (comparative evaluation).

5.3 Results

5.3.1 Evaluation of Likert scale items
The possible answers of the Likert scale were encoded with values ranging from 1 for “not interesting at all” to 4 for “very interesting”. Whereas the top 20 interests of profile type 1 scored $2.38 \pm 0.33$, the same selection of interest categories for profile type 2 scored higher with $2.80 \pm 0.39$. This trend could be found for all n-best selections (see table 5) reaching a maximum difference of 0.7 for the top 5 interests.

<table>
<thead>
<tr>
<th>Interest</th>
<th>Type 1 ($n=52$)</th>
<th>Type 2 ($n=52$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 5 interests</td>
<td>2.3808 0.3302</td>
<td>3.0840 0.4361</td>
</tr>
<tr>
<td>Top 10 interests</td>
<td>2.3923 0.3318</td>
<td>2.946 0.4166</td>
</tr>
<tr>
<td>Top 15 interests</td>
<td>2.3859 0.3309</td>
<td>2.8645 0.4050</td>
</tr>
<tr>
<td>Top 20 interests</td>
<td>2.3798 0.3300</td>
<td>2.8021 0.3963</td>
</tr>
</tbody>
</table>

The recommended Wikipedia articles (profile type recommendations) have been evaluated with an average score of $2.46 \pm 0.33$ by the participants.

5.3.2 Precision
The precision measures the ratio of relevant recommended items to all recommended items. For calculating the precision we considered items rated as “very interesting” and “interesting” as relevant to the user (true positive) and items rated as “hardly interesting” and “not interesting at all” were considered irrelevant (false positive). Figure 2 depicts the precision curves for different n-best selections of profile type 1 (red curve, dashed) and profile type 2 (blue curve, solid).

![Figure 2: Precision curves profile type 1 and 2](image)

Again profile type 2 (disambiguation heuristics applied) dominates profile type 1 (no disambiguation) in each n-best selection: Regarding the top 5 interests, users indicated a correct assignment for over 80%. For all inferred topics (top 20), at least two thirds are considered relevant by the users. Similarly, two of the top 3 Wikipedia articles recommended in the third step are considered relevant.

5.3.3 MAP and MRR
Mean Average Precision and Mean Reciprocal Rank answer the question of how well the interests are ranked at top-k and how early relevant results appear [13]. Again both MAP and MRR scored higher for profile type 2 (0.72 and 0.85) than for profile type 1 (0.50 and 0.68).

5.3.4 Comparative evaluation with a user’s friend
Profile type 4 was built up with top-10 interest categories of profile type 2 and 10 interest categories where no interest was assumed. The users were asked to evaluate this profile and send a link to a friend of theirs to do the same. The performance of our algorithm and the friend’s assessment was compared by the user’s evaluation (benchmark). Table 6 shows the confusion matrix comparing the performance of the algorithm introduced in this paper and the user’s friends (in brackets).

With a combined success rate of 74% versus 60% our approach clearly outperforms the friend in predicting the user’s interests. Differences in the above mean values are statistically significant ($p < 0.01$).
5.4 Analysis and Discussion

The results of the user study show that a user’s followees are not only a sufficiently broad basis for inferring interest profiles but these interest profiles are also a valid representation of the user’s interests. Profile type 2 scored better than profile type 1 in all quality measures calculated. This implies that the disambiguation heuristics have a significant impact on the quality of the generated interest profiles. With over 7 out of 10 items being relevant to the users our approach could achieve state of the art results and performed even better in predicting the users’ interests than their friends (thus humans) did.

6. PRODUCTION VERSUS CONSUMPTION BASED PROFILES

We did additional research to answer the introductory question, which also was raised by [17], of how interest profiles based on the user’s tweets (production) and followees (consumption) differ. Research on this questions also taps into the long-running debate on consumption vs. production online, where it is often argued that these two actions, particularly with regard to social media and digital content, are inseparable.

6.1 Approach

To provide the basis for a valid comparison of the two different profiles we generated them using the same approach and knowledge base, but extracted the entities from the user’s tweets in one case and from his followees in the other. However, we did not conduct a second user study to assess and compare the quality of the two different profile types, but only compare the created profiles.

In the first stage we extracted the entities for the consumption based profile of the user’s friend list as described in section 5 whereas for the production based profile the Illinois Wikifier [13], an external tool, was used to extract entities from the user’s tweets. In the subsequent stage, which was the same for both profile types regardless of their input, a spreading activation algorithm was performed on the Wikipedia Bitaxonomy [5] to aggregate the single interest entities to a more abstract and broader interest profile.

Finally, the interest profiles, which represent the user’s interests as a list of weighted Wikipedia categories are used to calculate the cosine similarity (as shown in eq. (4)) of the two different interest profiles.

\[
sim(X, Y) = \cos(\theta) = \frac{X \cdot Y}{||X||||Y||} \tag{4}
\]

where \(X\) is a vector representing the interest items’ weights of the production based profile and \(Y\) is the corresponding vector for the consumption based profile. As we use a taxonomy as knowledge base this allows us the comparison of the profiles on different levels of abstractness represented by varying parameters of the spreading activation algorithm (e.g. number of iterations or the decay factor).

6.2 Sample

To calculate the similarity of the two different profile types we selected a random sample of 50 twitter users from the Twitter sample endpoint, which allows to access a small sample of all public statuses and applied a set of selection criteria (e.g. only public accounts, a sufficient number of tweets and friends) on it to get suitable accounts for our experiment only. Apart from that we took the ten participants from the user study described in section 5 that rated the interest profile best and the ten that rated it worst. This leads to a total number of 70 twitter users for which both profile types and their cosine similarity were calculated.

6.3 Results

Compared to the followees of a user it appears to be easier to extract entities from the tweets as results showed that about 2.5 times more entities could be extracted using tweets rather than friends as input. Only about 2 % of the extracted entities (meaning Wikipedia pages) were shared of both sets. Whereas an intersection of 9 % could be found for the first level categories (representing the initially activated nodes). At the first glance these results indicate that the generated profiles and inferred interests do not seem to be too similar.

To gain a deeper insight the cosine similarity of the two interest profiles (list of interest categories and their corresponding weight) was calculated with a fixed decay factor (0.1, 0.2 and 0.3) and iterations ranging from 1 to 20. The number of iterations represents the different levels of abstraction, ranging from very concrete (one iteration meaning the initially activated nodes and their corresponding weight) to very abstract (20 iterations upwards in the Wikipedia
category taxonomy). As shown in figure 3, the cosine similarity of the concrete profiles (one iteration) comprising the weighted initially activated nodes is with about 0.66 higher as it would be expected given the small intersection of entities and categories. With an increasing number of iterations the similarity of both profiles is decreasing, whereby the low is reached later if the decay factor is smaller. As expected the similarity is rising for a high number of iterations again, as the categories reached by the spreading activation algorithm now are of a very abstract nature, i.e. the spreading activation accumulates in the top level categories. In general, profiles appear to be more similar on very concrete and abstract levels of the taxonomy.

Figure 4 compares the similarity of the 10 profiles that have been evaluated worst and the 10 that have been evaluated best in the user study that is described in section 5 above.

![Figure 4: Cosine similarity of production and consumption for 10 best and worst evaluated profiles](image)

The overall trends and the "u"-shape that can be seen in similarity with an increasing number of iterations are the same for both selections. However, the worst evaluated profiles have a higher similarity on a more concrete level (smaller number of iterations) whereas the best evaluated profiles seem to be more similar on an abstract level (higher number of iterations) approaching complete similarity at 20 iterations.

6.4 Analysis and Discussion

In general the cosine similarity of both profiles turns out to follow a "u"-shape along an increasing number of iterations. This means that profiles are more similar on very concrete and abstract levels of the Wikipedia Bitaxonomy (the used knowledge base). While the latter is expectable, since the spreading activation accumulates on the top level categories, the former is an interesting finding. In particular, since the profiles only share 2% of extracted entities. The larger intersection on the first level of categories (9%) is also not surprising, as the categories provide an abstraction of the very specific entities. That is, the number of (available) categories is smaller than the number of (available) entities. However, the high cosine similarity is an interesting finding, which suggests, that even though the intersection of entities and first level categories is low, the weights accumulate in the same categories. Still, the huge amount of categories in which the approaches differ seem to trigger different routes through the category graph to the top levels.

Even though there are differences in the concrete entities that could be extracted from the different sources of input, the information we get based on the consumption and production of the users’ twitter accounts are quite similar for certain levels of abstraction. The results therefore support the hypothesis that consumption and production online, particularly with regard to social media and digital content, are inseparable. Nevertheless there is further research needed to support the findings of this work and approach to an explanation for the significant differences in similarity on the different levels of abstractness.

7. CONCLUSION AND FUTURE WORK

In this paper we introduced an approach for inferring semantically meaningful interest profiles from the accounts a user follows on Twitter. Because the followees are the only input used, it is possible to generate interest profiles even for users that posted no tweets. Also, investigating the coverage of followee lists in terms of named entities in Wikipedia revealed that followees indeed provide sufficient input for creating comprehensive semantic interest profiles. As passive social media use is on the rise, the approach is an important contribution to the development of future social-media-based recommender systems that try to address the cold start problem. By conducting an extensive user study we could show that our approach achieved state of the art (and superhuman) results in predicting a user’s interests. A comparison of followee- and tweet-based profiles revealed high similarity on very concrete and abstract levels, suggesting that passive and active use are tightly coupled.

For future work we plan to extend our approach to other social networks such as Facebook (for which “likes” should be semantically equivalent to followees on Twitter). The evaluation showed that the disambiguation heuristics had a significant impact on the profile quality. Hence it appears to be promising use more sophisticated disambiguation and entity linking algorithms in the future and exploit recent advances in that field [21].

Further, we aim to investigate the relation between production and consumption based profiles in more detail. In particular, we are interested in whether a combination of both could improve the performance or if they can mutually benefit from each other.

8. ACKNOWLEDGMENTS

The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601.

9. REFERENCES


ABOUT THE AUTHORS:

Christoph Besel received a B.Sc. in Internet Computing from University of Passau, Germany in 2016. He is currently doing his Master’s in Web Science and Big Data Analytics at University College London, United Kingdom. His research interests include social media analytics, recommendation systems, applications of data mining and web economics in particular.

Jörg Schlötterer received his bachelor and master degree at the University of Passau, Germany in 2010 respectively 2013. He is currently pursuing his Ph.D. degree in Computer Science at the Professorship of Media Computer Science (University of Passau). His research interests center around information retrieval and connected topics, such as text mining, user profiling and search user interfaces.

Michael Granitzer has been Professor for Media Computer Science at University of Passau since 2012. Before, he was Scientific Director of the Know-Center Graz since 2010 and assistant professor at the Knowledge Management Institute of Graz University of Technology since 2008. In 2011, he was Marie Curie Research Fellow at Mendely Ltd. working on machine learning and information retrieval in academic knowledge bases. His research addresses topics in the field of Knowledge Discovery, Visual Analytics, Information Retrieval, Text Mining and Social Information Systems. He published over 180 mostly peer-reviewed publications and has been scientific coordinator and participant in several EU funded and nationally funded projects.
Using Mobile Messages to Improve Student Participation in Blended Courses: A Brazilian Case Study

Edgar Marçal, Rossana Andrade, Rosemeiry Melo, Windsor Viana, Eduardo Junqueira
Federal University of Ceará
Fortaleza-CE, Brazil
edgar@virtual.ufc.br, {rossana, rmelo}@ufc.br, {windson, eduardoj}@virtual.ufc.br

ABSTRACT
Studies have shown the benefits of mobile messages in education. They assist to fix the classroom lectures, help to remind study activities, and also allow the exchange of education messages among classmates. Researchers argue that more studies must be conducted to assess the extension of these benefits, and to better identify in which context mobile messaging is the best approach to be applied. In this sense, our study examined a post-graduate class consisting of 474 learners in which Short Message Service was used. The main goal was to find out if this technology would improve student participation in the course. Two groups of learners were created, control and experimental, in order to analyze the activities’ fulfillment rate and the response time. The results from the data analysis showed that students who received mobile messaging outperformed significantly students from the control group. The experimental group had better results both in completion rate and response time. They were almost 20% more effective at completing their assigned activity and up to 78 hours faster than students from the control group. In addition, a questionnaire was implemented in order to evaluate participants’ satisfaction with the use of SMS. Results found that most part of the students were satisfied and enjoyed the use of the mobile messages.

CCS Concepts
• Human-centered computing → Ubiquitous and mobile computing → Empirical studies in ubiquitous and mobile computing.

Keywords
Mobile Learning, Mobile Phone Messages, Blended Learning, Student Participation.

1. INTRODUCTION
Smartphones have changed the way people communicate. Multiple wireless communication methods are available in these devices, which turn them one of the main personal conversation tools. Several sectors of the economy have made use of mobile messaging to improve communication with their customers, highlighting the banks, e-commerce, and educational institutions [34].

By introducing mobile devices in education, the concept of Mobile Learning or m-learning emerges. It provides the sense of freedom to study anytime and anywhere with Internet access and the use of multimedia resources (e.g., audio, images, videos, etc.), which are associated with the playful aspect of learning. Mobile devices have been used in various areas of education, including science education for children [1, 2], applications for teaching high school mathematics [3, 4] and in a post-graduate computer applications course [5, 6].

Another approach emerges as an important educational model, Blended Learning, has shown significant results as: reduced dropout rates, improved student performances on examinations and increased collaboration among teachers and students [7, 8, 9]. Blended Learning combines traditional face-to-face (F2F) classroom learning practices with information and communication technologies. Researchers around the world have developed studies with Blended and Mobile Learning [10, 11, 12].

One of the main features of the mobile devices that have been integrated in blended courses is the short message service (SMS). Studies have shown the benefits of mobile messages in education, as fixing the content received in the classroom, reminder on what you need to study or do some work and exchange of education messages with classmates and the appropriate information is delivered in real time, allowing for just-in-time learning [13, 20, 24, 30]. Despite these positives, the researchers argue that more research should be conducted to evaluate the use of mobile messages in education, either because of the limitations of this technology that must be circumvented or to test them in different educational and regional contexts.

The aim of this paper is then to present a case study of the use of mobile messages texts as supplemental resources in a post-graduate blended course in Brazil. The main objective is to analyze the impact of m-learning in this context. In particular, we seek to know whether mobile messages can improve student participation and to comprehend the thoughts of those using this technology.

2. RELATED WORK
The Mobile Ecosystem Forum (MEF) conducted a study with approximately 6,000 participants from 9 countries in order to investigate the importance of mobile messaging for individuals and enterprises [34]. Figure 1 shows a graphic indicating the main institutions using mobile messaging to reach their customers. Furthermore, the graphic shows the messaging service used, if via Short Message Service or via a specific messaging application (chat app). For example, 33 per cent of people have received a text from their bank as opposed to 22 percent who have received a message from a chat app.
The expansion of the Internet messaging applications is dismissing the importance of the messaging service via SMS. However, studies justify the relevance of the use of SMS messages when you consider the communication between enterprises (e.g., companies, universities, government) and customers [37; 38; 39]. This mode of communication is known as A2P (application-to-person). According to the study from the MEF [34], institutions prefer to use SMS messages to chat app because they are more ubiquitous, since it not depend on specific application, operating system, or network connection.

Additionally, text messaging service is available in all mobile phones, unlike chat apps that must be installed by the users. In this case, both sender and receiver must have the same application. For example, if the teacher wants to send an activity by mobile phone, he must choose one (or more) chat application that students have installed on their mobile phones (e.g., Whatsapp, Facebook, Telegram).

To install the chat apps are required certain actions, such as performing specific settings, enter login information (username and password) and, depending on the mobile security specification, users must give permission to send and receive messages. In the case of SMS, the service is already factory-installed on the device. Thus, the user simply takes the mobile phone, put the carrier chip and may already receive SMS messages, without having to install applications or perform settings. This takes a neutral character to the SMS service compared with chat apps, as well as being easier for less experienced users with new technologies.

Adding the required expertise in the case of chat apps, another obstacle is the user actually needs to have Internet connection both to install the application as to exchange messages through it. This implies that the user must have a data plan with a mobile operator or rely on an Internet network via Wi-Fi. This restriction may make the use of chat apps in places where there is no support for mobile data network or intermittent connection.

In addition, other problems of chat apps are: privacy and expectation. Some applications such as Whatsapp, indicate whether the recipient receives the message and it was seen. On the one hand, the user may not want to share this information. On the other hand, this may generate an expectation it as to when the recipient will respond to your message. Moreover, Church and Oliveira present a study on mobile messaging that shows that most participants give more importance to the messages received by SMS that received via Whatsapp [37]. The authors claim that this is because SMS is a paid service and more mature.

The ability to receive and send messages with educational content is one of the key features of Mobile Learning paradigm (m-learning), which offer to teachers and students a more flexible learning approach. So, the objectives of m-learning are: improving resources for student learning; have access to lecture content anytime and anywhere; promote formal and non-formal learning; expand the boundaries of the classroom, and enable the development of innovative methods of teaching by using mobile computing technology [35].

Several studies show how these mobile technologies can enhance learning [27, 28, 29], for example: mobility to use the computing resources outside of traditional school environments, new opportunities to acquire content and encouraging the student to develop skills in real situations. Within the paradigm of m-learning, mobile messaging are used to increase the communication between teachers and students. Researches demonstrate the advantages of the use of mobile messages in education [11, 13, 14, 18, 19, 23].

A study on the impact of SMS among students enrolled in the second year of on-campus courses in New Zealand [13] used a system that allowed exchanges of information between students' mobile phones and a content server. The study concluded that the information transmitted must be condensed and relevant to students’ learning and that it should be based on some concrete activity. The study also recommended that information should reach the students directly, avoiding the need to access a web page to get the information. The high costs of the SMS transmissions were noted as a problem that inhibited interaction with the system developed to run the study. Petrova [13] says that SMS messages represent an attractive way to implement m-learning because (1) students can receive messages with educational content cost-free for them, (2) almost all types of mobile phones can handle this service, and (3) appropriate information is delivered in real time, allowing for just-in-time learning.

The area of learning languages has been one of the most widely used for research on the SMS messages in education [20, 21]. In the work [20] it was conducted an experiment for six months integrating SMS messages with face-to-face classes of English and

![Figure 1: Main areas using mobile messaging (via SMS and chat app) [34]](image-url)
French. Three types of text messages were sent in a schedule agreed upon with students: units of learning to read and store on mobiles not requiring student response; activities that ask questions or demand tasks; activities including collaborative tasks. The interaction between the students and the system was possible because the majority had contracts with the same provider, with free SMS between SIM cards on the same network. The authors point out that despite the small size, the short message service cans become a quick and effective tool for teaching languages. They conclude affirming that the research m-learning tends to become an important area in education.

In [24], the authors showed a study that aimed to investigate the impact of using SMS as learning support tool on students’ learning in an introductory programming course. The study examined the perceptions of 52 students of the advantages and disadvantages of the use of SMS as a learning support tool in their class. The educational content of the sent messages were developed and sequenced based on the analysis of the educational content of introductory programming course. Three types of educational contents were sent using SMS: review of programming concepts, hints to solve assignments and triggering questions. Students did not interact with the system. In the work [24], the analysis of the collected data showed that the use of SMS as learning support tool contributed significantly in improving students’ learning. All the interviewed students believed that the use of SMS technology as learning support tool has more advantages than disadvantages.

Other studies show the benefits of the use of instant messaging for education through the chat apps like WhatsApp and KakaoTalk [30, 31, 32, 36]. These studies highlight the messaging applications provide a virtual environment pleasant teaching that promotes the acquisition of profound knowledge through interaction with colleagues. However, problems are identified: security of information exchanged between users [31]; need for access to the Internet [32]; dispersion among students who use the application to talk with other people about other matters [30]; besides having to install the application on smartphone of the student (if he has a compatible) and know how to use it.

From the analysis of related studies cited in this section, we have identified the following benefits provided by the use of mobile messaging in education:

- Receiving mobile messages helps in fixing the content received in the classroom;
- Mobile messages are quick and easy to send and receive content;
- The mobile messages work as a service reminder on what you need to study or do some work;
- The use of short messaging service in mobile is familiar to most students;
- The ease of interaction stimulates the exchange of education messages with classmates;
- The questions sent via mobile messages keeps students connected to the course.

It can be seen that mobile messages are features that provide benefits to education. Despite the benefits previously mentioned, there are still barriers restricting the use of mobile messages on educational scenarios. Table 1 shows the problems encountered in the use of two types of messages and possible solutions, through the use of another transmission mode. The choice of which type of message mobile use will vary according to the situation and must take into account factors such as the existence of Internet connection at the place of use, users’ skills, type and size of messages, and availability of financial resources users. Depending on the context, a good strategy is also to combine these two approaches (SMS and chat app).

In some cases, the use of mobile messages does not show positive results in the improvement of teaching methods. For instance, in the work [21], it was implemented a comparative study on the use of cell phones versus computers to identify which technology is more suitable for answering exercises. Because of the effort to enter several texts on the mobile phone, most students reported that preferred to answer via computers. The authors claim that more research is needed to circumvent the limitations of mobile devices and identify what changes need to be made to content to better adapt it to that platform.

Mobile learning represents an extension of e-learning, while facilitating access to learning. The combination of e-learning and m-learning technologies in the classroom courses gives rise to the Blended Learning courses [15, 16, 25, 26]. Sherimon [14] define Blended Learning as the combination of traditional on-campus learning with the new technologies of online learning, like

| Table 1: Problems and solutions with the use of mobile messaging in education |
|-----------------------------------|-----------------------------------|
| **Problem with SMS messages** | **Solution via Chat Apps** |
| The limited amount of characters per SMS message is a restrictive factor. | Messages via Chat Apps have no limitation in number of characters. |
| The impossibility of the use of multimedia resources (such as images and audio) restricts the content type to text messages. | Messaging applications allow you to send multimedia messages like image, audio and video. |
| **Problem with Chat Apps** | **Solution via SMS** |
| Data connection need and concern for the security of data traffic on the Internet. | SMS messages do not need Internet connection to be transmitted. |
| Technical knowledge required for mobile application installation. | To send and receive SMS messages is not necessary to install an application on your phone. |
| Dispersion of the students and the teacher irritation due to lots of conversations of other subjects. | The SMS usage profile is more moderate and typically involves a smaller amount and more objective messages. |
Learning Management System (LMS) and mobile messages.

The existing literature about mobile learning points to a promising field of study. Particularly, regards to the assessment of its impact in different contexts, given the technology, economic and cultural differences in various parts of the planet. Along this perspective, a study with participants of different realities (urban and rural) using mobile messages is relevant, given that the contextual elements would enrich the panorama of studies already conducted.

3. **METHOD**

3.1 **Participants**

This research presents an experiment using mobile phones in a Blended Learning post-graduate course with 474 students at the Federal University of Ceará in Brazil. They were public school teachers engaged in the pedagogical coordinator role. The students were divided into 10 classes, each containing approximately 40 students. Table 2 summarizes the profile of the students.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>88% Female; 12% Male</td>
</tr>
<tr>
<td>Age group</td>
<td>67% over 40 years; 33% 25 - 40 years</td>
</tr>
<tr>
<td>Computer in home</td>
<td>91% Yes; 9% No</td>
</tr>
<tr>
<td>Broadband speed</td>
<td>59% Yes; 41% No</td>
</tr>
<tr>
<td>Access to the Internet</td>
<td>63% 1 to 3 times per week; 37% More than 3 times per week</td>
</tr>
<tr>
<td>E-learning experience</td>
<td>85% No; 15% Yes</td>
</tr>
</tbody>
</table>

From the profile it can be observed that the majority of participants were women over 40 years of age. During the course, also noted that most of them had children. So they had to take care of them, work and answer the activities while doing the course. Regarding access to the Internet, almost all students have a computer at home (91%), although, their time spent online was low (63% accessed at most three times a week). Moreover, many of them do not have broadband internet at home (41%). So, they access the online course in their work environment.

Participants had some knowledge of text editing software, which were enough to allow them completing the activities on the LMS. Most of them had no experience with online courses, only 15% of them have already done a course in this modality.

The classroom sessions of the course (face-to-face classes) were held in the following cities: Fortaleza, Ceará state capital and 2121 km away from the capital of Brazil; Caucaia; Icapui; Jijoca de Jericoacoara; Juazeiro do Norte; Quixeramobim; Santa Quiteria; Sobral; and Taua. Figure 2 shows a map indicating where the classes took place. Each city had one class with 40 students, except Fortaleza that had two classes. The state of Ceará has one of the lowest Human Development Index (HDI) levels in Brazil, registering at 0.723. However, this value represents an average among all cities in the state. If we consider only those cities where the study took place, the HDI falls to 0.689. The coordination of the course, from which the messages were sent, was in the city of Fortaleza. The average distance between coordination and other cities was 256 km, reaching up to 494 km away. Students attended class in nearby cities to their homes.

3.2 **Research Development**

The course was divided into ten disciplines, where a part of the course (20%) occurred in the traditional face-to-face (F2F) learning in sessions to present content in classroom. At these moments, there were presentations, practical activities and written assessments. Most of the course (80%) took place online, using the tools of the Learning Management System, such as forum, quiz and chat. Normally, a discipline lasted about two months and the students had to perform on average three virtual activities. The LMS was available 24 hours a day every day of the week for students can consult information or carry out the activities.

The SMS technology has been chosen for sending messages to students due to the conditions in which the course was held. In many places in the inner cities there was no Internet connection or it was very intermittent. Text messaging service is an affordable technology even in less sophisticated phones. Thus, we attempted to engage all students of the experimental group and limit the effects that financial hardship could have on research. Moreover, as were many students and they were distributed in several cities, it would be impossible to support and address the questions of the students about the installation and use of chat apps.
The students were divided into two groups. Group A, the control group, consisted of 374 students who did not receive mobile messages (SMS) and attended the course exclusively through the LMS. Group B, the experimental group, was made up of 100 students who received mobile messages and used the LMS. In each of the classes, 10 students were randomly selected to form Group B.

One of the disciplines of the course was selected to implement mobile learning. Three activities, each with a question, were proposed that students needed to respond to through the virtual environment on the Internet. These activities were made available in the LMS for all students and sent via SMS to the students in Group B. All students were required to answer the questions, using any text editor, and send the response via the LMS. Figure 3 shows a representation of the study and presents an example of one of the questions that was sent via SMS to the cell phones of students in the experimental group. In Figure 3, the message text is in English, but in the study the text was in Portuguese.

To achieve uniformity between the groups, the choice of participants was random and the significance level was 95%. The maximum difference between the sample proportion and the true population proportion (error) allowed was 10%. Thus, there was obtained an estimated size sample of 96 individuals which was approximate to 100 (experimental group), to allow replacement of any disparate data or invalid.

The study took place between the months of July and September 2012. Three mobile messages were sent, each with a question at an interval of 10 days between them. The mobile messages were sent to students' cell phones at the same time that the activity became available in the LMS. For each question, students had a deadline of 10 days to resolve and complete the activity. But answers after the deadline were also accepted.

Specifically in this study, we would like to know if the SMS messages increased the student participation. Thus, we sought answers to the following questions:

Q1: Does the use of mobile messaging encourage students to fulfill their assigned activities?
Q2: Does the use of mobile messaging make students to solve the questions more quickly?

4. RESULTS

4.1 Data Analysis

All data used in this study were obtained from the records contained in the LMS used in this course (Moodle). These records allow queries to be made about, for example, the exact instant that each student completed the activities. One of the basic assumptions to perform a statistical analysis is that the variables follow the Normal Distribution. In order to test the null hypothesis of normality of the variable, we used the Kolmogorov-Smirnov test\(^1\). The result of the test, for a 5% significance level, is that the null hypothesis was rejected. Then, the original data was submitted to the Box-Cox transformation \(^2\) in order to address the problem of non-normality. We used statistical tests for answering Q1 and Q2, and, as consequence, for verifying the results of the study.

474 students were divided into 10 classes. In each class, 10 students were randomly selected totaling 100 students to form the experimental group (Group B). Therefore, each class has a set of students’ part of the experimental group and the remaining part of the control group (Group A).

Compliance rate. The first point discussed was the difference between the average rates of compliance activities carried out by the control and experimental groups. We tried to identify if the number of students who completed the activities in the experimental group was higher than the number of those who completed the activities in the control group (Question Q1). The importance of this analysis is that the students who did not complete the activity did not receive grades, directly impacting their overall performance in the course.

To analyze the differences between the average rates of compliance, we postulated the following hypotheses:

H0: There is no difference between the rates of compliance between students in the experimental and control groups;
H1: The average rate of compliance in the control group is lower than that of the experimental group.

Taking compliance data from the activities of Groups A and B, we applied the z-test\(^2\) for Equality of rates, whose results can be found

\(^1\) The Kolmogorov-Smirnov test is a nonparametric test of the equality of continuous distributions that can be used to compare two samples.
\(^2\) Z-test follows the standard normal distribution under the null hypothesis.
in Table 3. From these results, we can conclude that, for a significance level of 5%, the hypothesis $H_0$ is rejected. We also can say that the average rate of compliance in Group A is smaller than in Group B (hypothesis $H_1$).

Table 3: Results for equality of rates of completion of the activity

<table>
<thead>
<tr>
<th></th>
<th>Group A (Control)</th>
<th>Group B (Experimental)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success (number)</td>
<td>299</td>
<td>93</td>
</tr>
<tr>
<td>Failure (number)</td>
<td>75</td>
<td>7</td>
</tr>
<tr>
<td>Success Rate</td>
<td>0.799</td>
<td>0.930</td>
</tr>
</tbody>
</table>

P-Value: 0.0011
Test Statistic Z: 3.065
Critical Z (0.05 one-tail): 1.645
Confidence Interval: 95%
Lower limit: 0.0661
Upper limit: 0.1949

Therefore, we can say with 95% confidence that the group that received SMS messages performed better in terms of completion of the activity than the group that did not receive SMS, at a rate of between 6.61% and 19.49%.

Response time. Another aspect analyzed was the time that students of different groups took to complete the activity. We sought to identify whether students who received SMS messages responded in a shorter time than students who did not (Question Q2). This analysis is important because questions can require a short time to be solved and thus require greater agility in the response of the students.

To analyze the differences between average response times, we postulated the following hypotheses:

$H_0$: The difference between the average response time among students of the experimental and control groups is equal to zero.

$H_1$: The difference between the average response time among students of the experimental and control groups is nonzero.

The results of the tests for the difference in average response time, presented in Table 4, indicate that for a significance level of 5%, the hypothesis $H_0$ is rejected. We can also conclude, with a 95% confidence level, that students in the group that received SMS messages responded more quickly, with a difference of between 1.97 and 78.9 hours.

To test the equality of variance assumption of two independent normal populations was used the t-test. As P-value is lower than 5%, reject the $H_0$ hypothesis, that is, the variances are not equal. The total number of observations presented in Table 4 (392) differs from the total number of students (474) because not all students completed the activity, and so it was not possible to include them in the analysis of response time.

Table 4: Results for homogeneity of variances (f-test$^3$) and for the average response time of the activity (t-test$^4$)

<table>
<thead>
<tr>
<th></th>
<th>Group A (Control)</th>
<th>Group B (Experimental)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive statistics (Non-normal distributions)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (hours)</td>
<td>185.171</td>
<td>144.730</td>
</tr>
<tr>
<td>Variance ($\sigma^2$)</td>
<td>32,089.116</td>
<td>11,170.044</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>299</td>
<td>93</td>
</tr>
<tr>
<td>Confidence Interval for mean: 95%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower limit</td>
<td>1.975</td>
<td></td>
</tr>
<tr>
<td>Upper limit</td>
<td>78.908</td>
<td></td>
</tr>
</tbody>
</table>

| Descriptive statistics (Normal distributions) |                  |                        |
| Mean (hours)   | 4.905             | 11.740                 |
| Variance ($\sigma^2$) | 0.713             | 12.462                 |
| Sample Size (n) | 299               | 93                     |
| Confidence Interval for mean: 95%                  |                  |                        |
| Lower limit    | -7.271            |                        |
| Upper limit    | -6.399            |                        |

Hypothesis F-test | Significance ($\alpha$) | 0.005
Degrees of freedom (numerator) | 298 | Degrees of freedom (denominator) | 92
P-value | 2,49151E-79 | F Statistics | 0,057
Standard error | 0,844 | Standard error | 3,530

Table 5 shows the results of the confidence intervals for the average response time. For this analysis, we took into account both the non-normalized and the normalized data. The graphs of Figure 4 and Figure 5 show that there is no intersection of the confidence intervals and sampling means of the two groups.

Figure 4: Confidence Interval for the average response time (Sample Non-normalized)

---

$^3$ The f-test has an F-distribution under the null hypothesis to test if two population variances are equal.
$^4$ The t-test uses statistical concepts to determine whether to reject a null hypothesis based on a Student’s t distribution.
Table 5: Confidence interval for the average response time of the activity

<table>
<thead>
<tr>
<th></th>
<th>Group A (Control)</th>
<th>Group B (Experimental)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-normal distributions</td>
<td></td>
</tr>
<tr>
<td>Mean (hours)</td>
<td>185.171</td>
<td>144.730</td>
</tr>
<tr>
<td>Variance ($\sigma^2$)</td>
<td>32089.116</td>
<td>11170.044</td>
</tr>
<tr>
<td>Confidence Interval for mean: 95%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower Bound</td>
<td>166.435</td>
<td>111.134</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>203.908</td>
<td>178.325</td>
</tr>
<tr>
<td></td>
<td>Normal distributions</td>
<td></td>
</tr>
<tr>
<td>Mean (hours)</td>
<td>4.905</td>
<td>11.740</td>
</tr>
<tr>
<td>Variance ($\sigma^2$)</td>
<td>0.713</td>
<td>12.462</td>
</tr>
<tr>
<td>Confidence Interval for mean: 95%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower Bound</td>
<td>4.693</td>
<td>11.360</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>5.117</td>
<td>12.121</td>
</tr>
</tbody>
</table>

Figure 5: Confidence Interval for the average response time (Sample normalized by Box-Cox transformation)

4.2 User Evaluation
To measure the level of acceptance of mobile messages in the course, we gave to the students in the experimental group an online survey after they completed their tasks. The questions of the questionnaire were as follows:

Q1 – Do you agree that sending messages is positive for the performance of activities?
Q2 – Did you feel more encouraged to answer the question after receiving the message on your mobile?
Q3 – Would you like the mobile messages to be used in other activities?
Q4 – Do you prefer to get the question on your phone rather than to have to consult the LMS?
Q5 – Did you have trouble reading the question received on your mobile?

Figure 6 summarizes the evaluation questionnaire responses. The objective was to determine the students’ opinion of the activity, whether they were satisfied with the use of mobile messages, and whether they would have liked to use this tool throughout the course. The questionnaire was composed of five propositions calling for respondents to indicate whether they agree or disagree.

From 100 students of the experimental group, 78 responded anonymously to the evaluation questionnaire. The results of this evaluation are encouraging. The majority (94%) of these students felt that mobile messages helped them complete the assigned tasks. Some 85% of students who received messages said that the messages stimulated them to action.

Approximately two-thirds of the respondents (67%) preferred mobile messages to activity in the virtual environment. Furthermore, almost all students who received the messages (92%) said they would like this feature to be used in other course activities. Finally, the survey also shows that mobile messaging was presented as an interesting tool to support teaching because of its relative ease of use; only 17% of students had difficulty reading the question that was sent to their cell phone. Levy says that the difficulty in using technology is one of the main factors leading students to drop out from e-learning courses [22].

In addition to the agree/disagree questions, the questionnaire provided a blank field in which course participants could express their opinions about the use of SMS messages in the course. Some written responses in this field can be grouped thematically:

Dispense access to Internet network
"It was great to receive messages via the phone, because I live far away from the city and do not have Internet access."

"It will help a lot, especially when you cannot access the LMS. As I already knew what the question was about, it was not necessary to consult the virtual environment to begin to answer it."

"I think it was a very interesting novelty. I confess that I felt very encouraged to use the phone because I can always be up to date with the activities, since the phone is with me 24 hours, other than the computer."
“It was very useful. I had no access because I was traveling and I was hard to get online and was quite helpful. I wish to continue receiving.”

“It was very nice to receive messages by phone, because I live too far from town and there is no Internet.”

**Works as a reminder to remember the activity**

“I don’t always have the time available to access the virtual environment and then I think the SMS messages are a way to reinforce our commitment to the course.”

“The initiative was great, because sometimes in the rush of daily work we cannot access the LMS, and we have our phone on hand and can receive messages wherever we are.”

**It helps people who have little time to study**

“I think this methodology and tool are very innovative and positive. Because even those who have computer with Internet at home, due to the everyday's rush, gets overwhelmed and slows activities.”

“I believe that mobile messaging is a way of reinforcing our commitment to the course and keep us watchful because not always the time at work allows we access the LMS.”

“Keep sending messages by mobile phone, because it is a great opportunity for people who are too busy.”

**Student appreciation**

“I would like this procedure to continue happening because I felt even more valued. After all, I had a question directed at me”

**Awareness of potential problems**

“I appreciate receiving questions via phone. However, I consider it important that the activity continue to be posted on the LMS to avoid potential problems with lack of cell phone service.”

“The mobile message helps but thinks that the computer offers better visibility.”

**5. DISCUSSION**

The questionnaire to evaluate participants’ satisfaction with the use of SMS found that most students were satisfied and enjoyed the use of new technology. The main benefits of mobile messages were that (1) they do not require the students (who are also teachers and are busy all day) to enter the LMS to visualize the current activities and (2) mobile messages arrive where students are and at any time, even in the absence of Internet access. These findings confer with the study of [19, 20], who concluded that the SMS enable students to learn in different contexts, regardless of location and time. Also in agreement with these results, 85% of students stated that SMS messages promoted a flexibility to study. When students were asked if SMS messages had positive effect on the performance of activities, 94% said yes. This finding agrees with the study of [11, 18], which concluded that most students believe that the use of SMS increases the interest in studies and has a positive influence on academic performance. [14] also claim that SMS messages have enormous educational potential to increase students’ motivation for distance learning courses, encouraging participation in the course and promoting greater interaction between them.

In most studies, the evaluation questionnaire has been the only way to identify the benefits of SMS in education. In this work, besides the questionnaire, an analysis of student participation in course was performed from the response times of the activities that were proposed to them. The results obtained with the statistical tests showed that the students who received mobile messages outperformed students who do not. The experimental group had better results both in completion rates and response times. They were almost 20% more effective in completing their assigned activities and up to 78 hours faster than students from the control group. The margin of confidence for these findings is 95%.

Another important result of this study is the fact that 92% of participants would like to continue receiving activities via SMS. This information agrees with the study of [11] which consisted that students have interest to continue to receive SMS messages, which are appropriate for learning. [11] also suggest that, before sending the messages, a survey was performed on the needs and preferences of students.

A disadvantage mentioned by some participants is that cell phone coverage is limited in some areas, making it difficult to receive messages. However, this is a technical issue and should be resolved when the telecom operators expand their coverage areas.

In this work, the SMS messages were sent in a single direction: Course for students. The main reason for choosing this approach was that students would have cost if they had to send messages. Some studies point the issue of costs with sending messages as an impediment to their use by students [14, 19]. [20] proposes, as a solution to this problem, groups of students who have the same operator, so they are not charged for the messages exchanged between them. However, such solutions depend on partnership and policy of the mobile phone operator.

Importantly, the results were achieved in a context where most of the students had no fixed or mobile connection to the Internet, or it was intermittent. In addition, the profile of participants was peculiar, teachers that were taking a pedagogical coordinator role in their schools. The fact that they are very busy and have little time to study may have favored the relevance of the SMS messages in the case study, since the push/notification characteristic of SMS messages made access to the content faster. These conditions may have biased the experiment results. So, we can't guarantee that the same results will be achieved in situations where students have access to fast and stable Internet connection. Further researches should be carried out to better evaluate the SMS in these conditions.

**6. CONCLUSION**

This study evaluated the use of mobile messages in a blended course for school teachers in Ceará, Brazil. It was concluded that the use of this technology increased student participation, as observed through rates of compliance, the times in which activities were completed and the questionnaire responses.

This study concluded that SMS-based m-learning can become useful tool in courses using the Blended Learning approach. Because cell phones are in such wide use, benefits can be realized even in economically disadvantaged regions. Moreover, from the testimony of participants, we can conclude that SMS messaging is a relevant alternative to reinforcement learning for those who work...
all day. Participants’ testimonies highlighted the importance of receiving the summary of activities during the workday.

The contribution of this study becomes more important when considering that the SMS does not need internet and sophisticated cell phones to work. This increases the scope of the solution, because it allows you to include students who do not have easy access to broadband Internet, which is a reality in cities in the interior. Furthermore, the technologies of wireless Internet high speed (such as 3G and 4G) did not arrive at these locations. This situation complicates the use of instant messaging (via internet) apps like Whatsapp and reinforces the relevance of this solution (via SMS).

Regarding future work, this study raises some interesting technological and pedagogical issues. In this course, SMS messages were designed to be used in interventions for short periods of time. We believe that if mobile messages are used over a long period of time, the benefits from its use for learning motivation will increase.

Another promising application is the use of servers for sending and receiving messages, although this practice requires considerable costs for deployment and maintenance. From the pedagogical point of view, it is worth investigating whether mobile messages can lead to gains in students’ learning, not just in increasing their participation. To answer this question, it will be necessary to do further research in the area of educational assessment.

7. ACKNOWLEDGEMENTS
Thanks to CNPq for The financial support (Research scholarship) granted to Rossana M. C Andrade.

8. REFERENCES


ABOUT THE AUTHORS:

Edgar Marçal de Barros Filho received his PhD degree in Computer Science from the Federal University of Ceará (UFC) and a master degree from the UFC (2005). He is an Adjunct Professor at the Federal University of Ceará, crowded at Virtual University Institute. He has experience in information technology project management area in education, commercial project management and systems analysis, acting on the following topics: mobile computing, mobile learning, systems analysis, software engineering and distance education via the Internet.

Rossana Maria de Castro Andrade is the founder of the Group of Computer Networks, Software Engineering, and Systems (GREat) of the Federal University of Ceará (UFC), Brazil. For 15 (fifteen) years she has been working on R&D with telecommunications companies in the development of mobile applications and software in general. Nowadays, she is the coordinator of the UFC Graduate Program (master and PhD) in Computer Science. She has PhD in Computer Science at Ottawa University, Canada (2001) and master and bachelor degree in Computer Science in Brazil, respectively, at the Federal University of Paraíba (1992) and State University of Ceará (1989). Rossana Andrade has experience in research, development, and innovation in the areas on Computer Science and Telecommunications, acting specially in software engineering and computer networks.

Windsom Viana de Carvalho received his doctorate (spécialité Informatique) from the Université de Grenoble, France (2010), and a master degree in Computer Science from the Federal University of Ceará (UFC) (2005). He is an Adjunct Professor at the Federal University of Ceará, crowded at Virtual University Institute. He has experience in computer science, with emphasis on Mobile Computing, Multimedia and Software Engineering, acting on the following topics: mobile and ubiquitous computing, context awareness, middleware, ICT applied to aid teaching and learning, assistive technologies, management of multimedia documents, pervasive games and recommender systems.

Eduardo Santos Junqueira Rodrigues received his PhD degree in Education from Michigan State University. He is an Adjunct Professor at the Federal University of Ceará, crowded at Virtual University Institute. Dr. Junqueira has experience in the areas of Education and Social Communication, with an emphasis on new literacies and multiliteracies, digital inclusion, multimodality and hypertext, ethnoarchaeological research and online education. Currently he coordinates studies about hipermodal navigation in learning contexts using the eye tracking technique. He is the leader of the research group Languages and Networked Education. WebSite: http://www2.virtual.ufc.br/ler/

Rosemeiry Melo earned a Master’s degree in Rural Economics and a PhD degree in Economics. She teaches in undergraduate courses in Agronomy, Animal Husbandry and Fisheries Engineering at the Federal University of Ceará. Her areas of interest are Economics of Production, International Trade, and Economic Development.
Discriminating Graph Pattern Mining from Gene Expression Data

Fabio Fassetti  
DIMES, University of Calabria  
87036 Rende (CS), Italy  
f.fassetti@dimes.unical.it

Simona E. Rombo  
DMI, University of Palermo  
90123 Palermo (PA), Italy  
simona.rombo@unipa.it

Cristina Serrao  
DIMES, University of Calabria  
87036 Rende (CS), Italy  
c.serrao@dimes.unical.it

ABSTRACT

We consider the problem of mining gene expression data in order to single out interesting features that characterize healthy/unhealthy samples of an input dataset. We present an approach based on a network model of the input gene expression data, where there is a labelled graph for each sample. To the best of our knowledge, this is the first attempt to build a different graph for each sample and, then, to have a database of graphs for representing a sample set. Our main goal is that of singling out interesting differences among graphs belonging to the two different sample sets. Differently from the other approaches presented in the literature, our techniques is able to take into account important local similarities, and also collaborative effects involving interactions between multiple genes. In particular, we use edge-labelled graphs and we measure the discriminating power of a pattern based on such edge weights, which are representative of how much relevant is the co-expression between two genes.

CCS Concepts

• Computing methodologies → Motif discovery;
• Applied computing → Bioinformatics;

Keywords

Biological Networks, Gene expression data, Pattern Mining.

1. INTRODUCTION

Mechanisms regulating the organization and functioning of cells are still not completely understood, although it is commonly recognized that they are based on the interplay of several different factors. For many decades single molecules playing important roles in the cell, such as proteins, genes and RNA, have been deeply studied as independent objects. At the beginning of this century, after that the genome sequencing of many organisms, among which human, have been completed, the attention has turned on how cellular reactions. Building all these kinds of networks is possible thanks to the information stored in public interaction databases and mainly obtained by high-throughput technologies. Here we focus on networks based on gene expression data, and we recall some basic notions below on this type of data (the interested reader can find exhaustive information on the use of genome-wide gene expression data at [21]).

The transcriptome of a cell comprises mRNA, tRNA, rRNA, and short regulatory RNAs. In order to generate large-scale gene expression data, biologists use microarray experiments, that is, they measure genome-wide gene expression levels of mRNA in a cell or a tissue sample under a particular condition. A microarray chip quantifies the hybridization of fluorescent labelled target nucleotide sequences to define complementary probe sequences that are spotted on a glass or silicon slide. More details about microarrays may be find in [4, 19, 7].

In the last few years, also more sophisticated techniques have been developed such as next generation sequencing (RNA-seq) [13, 25]. RNA-seq has a wide variety of applications such as the measurement of gene expression levels from transcribed mRNA sequences.

All these technologies have revolutionized the biological research but it is challenging to interpret the direct results from experiments to investigate complex biological mechanisms. Therefore a number of techniques have been proposed to switch from a tabular to a network representation (see [20] to have some references about them).
Traditional techniques start from microarray measurements to find out the expression level of each gene in each of the samples analysed; this data is used to define the so called “profile expression” of each gene over the sample-set. Statistical, machine learning or soft-computing techniques have been introduced for the co-expression networks construction, but all of them need to look at the sample-set globally.

Actually, it has been observed that the expression profiles often share local rather than global similarities [20], so if one tries to model cellular mechanisms of an organism through a graph, some potentially powerful details of each interaction may be left aside.

Here we consider the problem of identifying interesting differences between two input sample sets, associated to healthy and unhealthy individuals, respectively. To this aim, we propose an approach based on two main characteristics: (i) a representation of gene co-expression data able to take into account local similarities, and (ii) the definition of a suitable notion of pattern useful to capture the differences between the two input sample sets. In particular, our model emphasizes the importance of locality by turning the microarray dataset into a graph dataset, where there is a labelled graph for each sample. Note that, in this context, the number of samples is much smaller than the number of genes. The main aim of our approach is that of singling out interesting differences between healthy and unhealthy samples, through the extraction of “discriminating patterns” among graphs belonging to the two different sample sets.

It is worth to point out that common discriminating graph pattern mining approaches have been shown to achieve great success by mining the graph patterns that occur with disproportionate frequency in some classes versus others [28]. However, this kind of information may be not enough when mining biological graph patterns, especially if one wants to capture those interactions that can be related to a certain pathological phenotype. Indeed, diseases such as cancer are often related to collateral effects involving interactions between multiple genes or proteins [5, 27]. Therefore the discriminating power of a pattern should be higher than the one of all its sub-patterns. We contribute in this direction both by enumerating patterns with node labels, which are associated to how much relevant is the co-expression between two genes, and by introducing a measure of how much discriminating is a pattern, based on the edge weights.

The rest of the paper is organized as follows. Section 2 surveys the work mostly related with ours. Section 3 introduces preliminary concepts and notation. Section 3.3 is devoted to the building of the networks which represent the feed of our technique. Section 4 presents the main problem we are interested in solving. Section 5 describes the contrast graph pattern mining algorithm to tackle the problem at hand. Section 6 reports the experimental campaign we conduct to validate our technique. Finally, Section 7 draws the conclusions.

2. RELATED WORK
Several approaches have been proposed in the last few years in order to infer information from biological networks. They are mainly based on alignment techniques (see [6, 16] for useful reviews), clustering (see [3, 18]) and pattern extraction (e.g., [11, 26]). Most of such approaches aim at finding evolutive conservations among different organisms, extracting functional modules or they can be used to predict the biological function of some not yet well characterized cellular components. Recently, the attention is turning on better understanding how the interactions among cellular components may influence the emergence of diseases in complex organisms, such as human. In this case, genotype to phenotype relationships, as well as those genomic variations involved in human disease, inherited in the germ line or acquired systematically, are suitably mapped [24]. In the remaining part of this section, we will focus our attention on those approaches based on suitable notions of patterns useful to distinguish different classes of biological graph samples, since they are the most related to the research work presented here.

In [23] the notion of minimal contrast subgraph pattern is introduced in order to single out structural differences between two collections of graphs. The approach showed to be useful in chemical compound comparison and building of graph classification models. Improved algorithms for the extraction of this kind of patterns were proposed in [29].

The authors of [28] consider the bioassay records for anti-cancer screen tests with different cancer cell lines and they build datasets belonging to a certain type of cancer screen with the outcome active or inactive. They propose an approach to distinguish these two classes based on the search of dissimilar graph patterns, according to a mining framework exploiting the correlation between structural similarity and significance similarity.

Synergy graph patterns have been defined in [26] by referring to sub-graphs such that the relationships among the nodes are highly inseparable. The authors of [26] use confidence values to calculate discriminating power scores of graph patterns and consider only those graph patterns whose discriminating powers are much higher than all their sub-graphs. They apply a classification algorithm based on synergy graph patterns to real-life datasets such as a AIDS antiviral screen chemical compounds dataset and anti-cancer screen datasets.

As the approaches summarized above, we define a notion of graph pattern able to capture significative differences between two input graph sample sets. The main differences with respect to the other techniques are that, first of all, we enumerate graph patterns by considering the unequeness of node labels. This allows to consider how much meaningful is the co-expression between two genes. Furthermore, while the discriminating power of the other patterns recalled above is mainly based on their support, the discriminating measure we introduce is able to take into account also the edge weights. This way we can capture local similarities that, as discussed in the Introduction, are important to single out biologically meaningful differences.

3. NETWORK MODEL
The basic input data considered here is gene expression data. Gene expression data contains information about the expression level of several genes on the set of analysed individuals. They are represented as a multiset of tuples on a set of at-
Given a population \( p \) of \( n \) genes, we first formally introduce a graph \((V,E)\) where \( V \) is a set of nodes (or vertices), each identified by a unique label, and \( E \) is a set of edges, i.e., unordered pairs \((v,w)\) where \( v, w \in V \). A sequence \( v_1, \ldots, v_h \) of nodes in \( V \) such that \((v_i, v_{i+1}) \in E \) for any \( i \in 1..(h-1) \) is a path from \( v_1 \) to \( v_h \). We recall that a graph is connected if for each pair of nodes \( v \) and \( w \), there is at least a path from \( v \) to \( w \).

Given an input set of gene expression data, in our model each gene is associated to a node \( v \) of a labeled undirected graph. Each edge connecting a pair of genes has two different weights: (1) the strength of the relationship between these two genes, and (2) the relevance. The latter is useful to make the analysis more robust to statistical fluctuations, since, roughly speaking, it represents the exceptionality of the strength associated with \( e \), w.r.t. the expected value.

In the following of the section, we first formally introduce the notions of strength and relevance and, then, we formalize the adopted network model.

### 3.1 Strength of the relationships

Given a population \( D \), an individual \( t \) of \( D \) and two genes \( a_i \) and \( a_j \), we aim at relating the pair of genes \( a_i \) and \( a_j \) for \( t \) to the correlation \( \rho_{ij} \) between \( t(a_i) \) and \( t(a_j) \). A first problem is to estimate the correlation between \( t(a_i) \) and \( t(a_j) \), for each tuple \( t \) and for any pair of attributes \( a_i \) and \( a_j \). Note that, despite classical approaches, the correlation considered here is based on only two observations.

Let \( X_i^t \) (\( X_j^t \), resp.) be the random variables associated with \( t(a_i) \) (\( t(a_j) \), resp.), and consider the bivariate normal distribution having mean vector \( \mu_{ij} \) and covariance matrix \( \Sigma_{ij} \), where:

\[
\mu_{ij} = \begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix}, \quad \Sigma_{ij} = \begin{pmatrix} \sigma_i^2 & \rho_{ij} \sigma_i \sigma_j \\ \rho_{ij} \sigma_i \sigma_j & \sigma_j^2 \end{pmatrix},
\]

\( \mu_i \) (\( \mu_j \), resp.) is the mean value of attribute \( a_i \) (\( a_j \), resp.), \( \sigma_i \) (\( \sigma_j \), resp.) is the standard deviation of attribute \( a_i \) (\( a_j \), resp.), then independent from \( t \), and \( \rho_{ij} \) is the correlation between \( X_i^t \) and \( X_j^t \).

In order to emphasize the impact of the observed values, an interesting value of correlation \( \rho_{ij}^* \) between \( X_i^t \) and \( X_j^t \) can be estimated by inferring the value \( \rho_{ij}^* \) maximizing the probability of observing the two-dimensional point \((t(a_i), t(a_j))\) and, thus, it can be suitably employed as strength of the relationship between them.

**Definition 1 (Strength).** Given a population \( D \), an individual \( t \) in \( D \) and two genes \( a_i \) and \( a_j \), the strength of the relation between \( a_i \) and \( a_j \) for \( t \) is the value of correlation that maximizes the probability of observing \( t(a_i) \) and \( t(a_j) \).

Since there is a one-to-one correspondence between an individual and its representing tuple, for the sake of simplicity, we employ the same symbol \( t \) to denote both the individual and its corresponding tuple in the dataset.

**Strength computation**

In order to ease the computation, we normalize each value according to the mean and the standard deviation of the associated attribute. Hence, we compute \( \hat{t}_i = \frac{t(a_i) - \mu_i}{\sigma_i} \) and \( \hat{t}_j = \frac{t(a_j) - \mu_j}{\sigma_j} \).

Let \( X_i^t \) (\( X_j^t \), resp.) be the random variable associated with \( \hat{t}_i \) (\( \hat{t}_j \), resp.) and consider the bivariate normal distribution with components \( X_i^t \) and \( X_j^t \), mean vector \( \hat{\mu} \) and covariance matrix \( \hat{\Sigma} \), where:

\[
\hat{\mu}_{ij} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \hat{\Sigma}_{ij} = \begin{pmatrix} 1 & \rho_{ij} \hat{t}_i \hat{t}_j \\ \rho_{ij} \hat{t}_i \hat{t}_j & 1 \end{pmatrix}.
\]

Thus, the bivariate normal distribution can be written as:

\[
f(x, y, \rho_{ij}) = \frac{1}{2\pi\sqrt{1 - \rho_{ij}^2}} \cdot e^{-\frac{1}{2(1-\rho_{ij}^2)}(x^2 + y^2 - 2\rho_{ij}xy)}.
\]

The aim is to find the value \( \hat{p}_{ij} \) of \( \rho_{ij} \) such that the value of \( f \) in the point \((\hat{t}_i, \hat{t}_j, \rho_{ij})\) is maximum, in the formula:

\[
\hat{p}_{ij} = \arg \max_{\rho_{ij}} f(\hat{t}_i, \hat{t}_j, \rho_{ij}),
\]

which represents the strength between genes \( a_i \) and \( a_j \) for \( t \). It can be proved that the stationary points are obtainable as the solution of

\[
\rho^3 - \rho^2 \hat{t}_i \hat{t}_j + \rho (\hat{t}_i^2 + \hat{t}_j^2 - 1) - \hat{t}_i \hat{t}_j = 0. \tag{1}
\]

### 3.2 Relevance of the relationships

In the previous section, we estimate how much two observations are correlated. In order to make this estimation more robust, we measure the probability that a possible high value of correlation is not due by chance. The underlying idea is to test the null hypothesis under which a high value of correlation could be implied by a certain value of expression of a gene for an individual. In other words, given a certain value of expression, it could be quite high the probability that the level of expression of another gene leads to a high value of correlation.

**Definition 2 (Relevance).** Given a population \( D \), an individual \( t \) of \( D \) and two genes \( a_i \) and \( a_j \), the relevance of the relation between \( a_i \) and \( a_j \) for \( t \) is the minimum between the probability of observing a strength smaller than \( \hat{p}_{ij} \) given the level of expression of \( a_i \) equal to \( t(a_i) \) and the probability of observing a strength smaller than \( \hat{p}_{ij} \) given the level of expression of \( a_j \) equal to \( t(a_j) \).

Intuitively speaking, the higher the relevance the smaller is the probability that the observed value of correlation is due by chance.

**Relevance computation**

Let \( t_i^* \) (\( t_j^* \), resp.) be the observed expression level of gene \( a_i \) (\( a_j \), resp.) in \( t \) and let \( \rho^* \) be the strength associated with \( \rho^* \).

---

[1]The reader is referred to the Appendix A. for the details.
such that the strength of $t$ and $t_0$ between is a function associating each edge in $E$ then obtain a database of $m$ $(P_i, E, \eta, \pi)$, where the $SR$-network $N_i$ is associated with the $i$-th individual of $DS$.

4. Statement of the Problem

This section is devoted to formally introduce the main problem we are interested in solving.

The reader is referred to the Appendix B. for the details.

Given a population $DS$, suppose that it is a-priori partitioned in two groups $DS_1, DS_2$ on the basis of certain properties of the samples (i.e., healthy vs unhealthy).

The goal is to single out peculiarities of a sub-population w.r.t. the other one. This can be exploited to shedding light on the characteristics that distinguish individuals of $DS_1$ from individuals of $DS_2$. Since this kind of knowledge is often related to collaborative effects involving interactions between multiple genes or proteins [5] [27], we aim at mining graph patterns that can be in charge of the separation between the sub-populations at hand. In particular, we search for patterns that are representative of one sub-population, but not of the other one.

First, we introduce the notion of pattern which is the building block of the knowledge we want to mine.

**Definition 4 (Pattern).** Given a SR-networks database $N$ defined on a set of nodes $V$, a pattern $P$ in $N$ is a connected graph $(V, P, E, \rho)$ with $V \subseteq V$.

A pattern $P' = (V_{P'}, E_{P'})$ is a sub-pattern of $P = (V_P, E_P)$ (or, equivalently $P$ is a super-pattern of $P'$) if $V_{P'} \subseteq V_P$ and $E_{P'} \subseteq E_P$.

Given a $SR$-network $N = (V, E, \eta, \pi)$ and a pattern $P = (V_P, E_P)$, there is a match of $P$ in $N$ if and only if $V_P \subseteq V$ and $E_P \subseteq E$.

Since, by construction, all networks and all patterns are defined on the same set of nodes and each node is different from each other being associated with a different gene, the following property clearly holds.

**Property 1.** Given a database $N$ and a pattern $P$ in $N$, for any network $N'$ in $N$, there exists only one match of $P$ in $N'$.

Therefore we say that a network $N$ matches a pattern $P$ (or, $P$ occurs in $N$), meaning that there is a match of $P$ in $N$, and such a match is unambiguously determined.

Next, we extend the notion of strength and relevance provided in Section 3.1 and 3.2 to a pattern $P$ in an $SR$-network $N = (V, E, \eta, \pi)$. In particular, the strength $\eta(P, N)$ of the match of a pattern $P = (V_P, E_P)$ in $N$ is defined as:

$$\eta(P, N) = \frac{1}{|E_P|} \sum_{e \in E_P} \eta(e)$$

while the relevance of $P$ in $N$ is defined as the product of the relevancies of the edges of $P$ in $N$:

$$\pi(P, N) = \prod_{e \in E_P} \pi(e)$$

The ratio underlying these definitions is that, in order for a pattern to be relevant, all its edges should be relevant. Since the relevance is a sort of “probability of observing the edge”, the product of the relevancies, similarly to the probability of intersection between events, takes into account this issue.

In order to evaluate how much a pattern is representative of a (sub)population, we need to compute how much a pattern is common to occur in that population. Note that simply counting the number of $SR$-networks in the database matching the pattern could be misleading, since all the information...
coming from strength and relevance is neglected. To obtain a robust measure, we evaluate the commonness of a patterns \( P \) in a (sub)population \( N \) and denote it by \( s(P, N) \), where only relevant matches are considered, and the strengths of the matches are summed:

\[
s(P, N) = \sum_{N' \in \mathcal{N} : s(P, N') > r_r} \eta(P, N').
\]

According to the statistical test theory, \( r_r \) can be set to the standard values for testing hypotheses.

**Property 2.** Given a database \( N \) and a pattern \( P \), the commonness of \( P \) in \( N \) is upper bounded by the support of \( P \) in \( N \), i.e., the number of SR-networks in \( N \) where \( P \) occurs.

The property immediately follows from the observations that the strength of a pattern in a SR-network \( N \) ranges in \([0, 1]\) and that it is 0 if the pattern has no matches in \( N \), while the support is 0 if the pattern has no matches in \( N \) and it is 1 if the pattern occurs in \( N \).

### 4.1 Discriminating Pattern

In order to measure the discriminating power of a pattern we resort to the notion of information gain [14] and adapt it to our context. The aim is to measure the change in information entropy [10] led by the pattern.

Let \( N \) be a population partitioned in two sub-populations \( N_1 \) and \( N_2 \) and let \( P \) be a pattern. The **discriminating power** of \( P \), denoted as \( pow(P) \), is the gain in entropy

\[
pow(P) = H(N) - H(N|P)
\]

namely, the difference between the entropy of the population \( H(N) \) and the entropy \( H(N|P) \) of the population given the pattern.

Thus, in order to define the discriminating power, we need to adapt the notion of entropy to our context.

The **information entropy** \( H(N) \) is:

\[
H(N) = -\frac{|N_1|}{|N|} \log \frac{|N_1|}{|N|} - \frac{|N_2|}{|N|} \log \frac{|N_2|}{|N|}.
\]

As for the information entropy conditioned by the pattern \( P \), \( H(N|P) \), we note that the pattern \( P \) partitions the population \( N \) in two groups of individuals those where \( P \) is relevant, denoted as \( N^P \), and those where the pattern \( P \) it not relevant, denoted as \( N^{P'} \). The entropy of \( N \) conditioned by \( P \) can be, then, computed as:

\[
H(N|P) = H(N^P) \cdot q + H(N^{P'}) \cdot (1 - q)
\]

with:

\[
q = \frac{s(P, N_1) + s(P, N_2)}{|N|}, \quad \text{and:}
\]

\[
H(N^P) = -q_1 \log q_1 - (1 - q_1) \log(1 - q_1)
\]

\[
H(N^{P'}) = -q_2 \log q_2 - (1 - q_2) \log(1 - q_2)
\]

with:

\[
q_1 = \frac{s(P, N_1)}{s(P, N_1) + s(P, N_2)} \quad \text{and}
\]

\[
q_2 = \frac{|N_1| - s(P, N_1)}{|N_1| - s(P, N_1) + |N_2| - s(P, N_2)}.
\]

Next, we provide a formal definition of the particular type of patterns we are interested in.

The patterns that the algorithm must highlight are those ones that give more discriminating than their sub-graphs (i.e. the discriminating power of the pattern should be higher than the one of all its sub-patterns). Moreover, if two patterns have the same discriminating power we take only the one with maximum commonness. Indeed, the discriminating power of a pattern that is low-supported can harm the accuracy of the analysis due to overfitting, then also in this case we need to take into account the commonness.

**Definition 5 (Discriminating pattern).** A pattern \( P \) is a discriminating if and only if for each pattern \( P' \) sub-pattern of \( P \) either \( pow(P) > pow(P') \) or \( pow(P) = pow(P') \) and \( s(P) > s(P') \).

This definition is relevant in biological fields as disease such as cancer are often related to collaborative effects involving interactions between multiple genes or proteins [5] [27], so those patterns whose discriminating power gives us more information than the one of their sub-patterns seemed to be really interesting, as it means that the relations among genes in the pattern can give more information about the sub-population than their sub-sets can do.

### 4.2 Problem definition

Patterns generated according to this definition may be redundant. Indeed, suppose you have mined a pattern \( P \) which is discriminating since its discriminating power is higher than that of its sub-patterns. If there is a pattern \( P' \) super-pattern \( P \) such that \( pow(P') > pow(P) \), according to Definition 5, \( P' \) is a discriminating patterns as well. To avoid keeping both \( P \) and \( P' \) in the result set of discriminating patterns, we resort to a notion of maximality as formalized in the following definition.

**Definition 6 (Maximal discriminating pattern).** A discriminating pattern \( P \) is said to be maximal if and only if there is not a discriminating pattern \( P' \) such that \( P' \) is a sub-pattern of \( P \).

Thus, the problem we are interested in solving consists in singling out all the maximal discriminating patterns.

### 5. ALGORITHM

This Section is devoted at presenting the algorithm we proposed to mine discriminating patterns, sketched in Algorithm 1.

We now present the algorithm proposed to mine discriminating patterns, sketched in Algorithm 1. The algorithm, basically, visits in depth and tries to prune the search space consisting in connected subgraphs of the input networks. In more details, for each subpopulation we build the associated database of SR-networks as described in Section 3. Next, we mine patterns in two phases by singling out patterns over represented in subpopulation \( DS_1 \) against \( DS_2 \) and then those over represented in subpopulation \( DS_2 \) against \( DS_1 \).
Algorithm 1: Discovering Discriminating Pattern

<table>
<thead>
<tr>
<th>Input:</th>
<th>Gene expression data DS (partitioned in DS1 and in DS2), thresholds ( \tau_s ) and ( \tau_r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>Discriminating patterns, res</td>
</tr>
<tr>
<td>N1 ← BUILDSRNETWORKS(DS1, ( \tau_r, \tau_s ))</td>
<td></td>
</tr>
<tr>
<td>N2 ← BUILDSRNETWORKS(DS2, ( \tau_r, \tau_s ))</td>
<td></td>
</tr>
<tr>
<td>foreach ( N_{\text{MAX}} \in { N_1, N_2 } ) do</td>
<td></td>
</tr>
<tr>
<td>edges ← SORTEDGES(( N_{\text{MAX}} ))</td>
<td></td>
</tr>
<tr>
<td>res ← PATTERNMINE(( \emptyset, ) edges, ( \emptyset, ) ( \tau_r, ) ( \tau_s ) \cdot (</td>
<td>N_{\text{MAX}}</td>
</tr>
<tr>
<td>Delete non-maximal pattern in res</td>
<td></td>
</tr>
<tr>
<td>return res</td>
<td></td>
</tr>
</tbody>
</table>

Function PATTERNMINE(\( P_{\text{cur}}, \) neighs, visitedEdges)

<table>
<thead>
<tr>
<th>Input:</th>
<th>Current Pattern ( P_{\text{cur}} ), pattern neighborhood [neighs] already visited edges [( vE )] thresholds ( \tau_r ) and ( T_s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>Discriminating pattern candidates [res]</td>
</tr>
<tr>
<td>foreach ( e ) in neighs do</td>
<td></td>
</tr>
<tr>
<td>( P_{\text{next}} \leftarrow P_{\text{cur}} \cup {e} )</td>
<td></td>
</tr>
<tr>
<td>( vE \leftarrow vE \cup {e} )</td>
<td></td>
</tr>
<tr>
<td>neighs( <em>{\text{next}} \leftarrow ) COMPUTEНЕIGHBORS(( P</em>{\text{next}} ))</td>
<td></td>
</tr>
<tr>
<td>( c \leftarrow ) COMPUTECOMMONNESS(( P_{\text{next}}, \tau_r ))</td>
<td></td>
</tr>
<tr>
<td>if ( c &gt; T_s ) then</td>
<td></td>
</tr>
<tr>
<td>if ISDISCRIMINATING(( P_{\text{next}}, c, ) res) then</td>
<td></td>
</tr>
<tr>
<td></td>
<td>res ← res ( \cup P_{\text{next}} )</td>
</tr>
<tr>
<td>if neighs( _{\text{next}} \neq \emptyset ) then</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PATTERNMINE(( P_{\text{next}}, ) neighs( _{\text{next}}, ) vE, ( \tau_r, T_s ))</td>
</tr>
<tr>
<td>else</td>
<td></td>
</tr>
<tr>
<td></td>
<td>if neighs( <em>{\text{next}} \neq \emptyset ) and CANRISE(( P</em>{\text{next}} )) then</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In other words, we aim at detecting patterns which are common (the commonness is high) in one subpopulation and rare (the commonness is low) in the other one.

As a second step, edges are sorted according to their average strength over the set of SR-networks, in order to find firstly patterns more representative and then potentially more interesting. Moreover, such a sorting is likely to make more effective succeeding pruning rules. Note that such a sorting change during the analysis of the search space since for each pattern under consideration, the set of networks in which the pattern is relevant changes.

Next, the function PATTERNMINE, which is the core of the algorithm, is called. It receives the current pattern \( P_{\text{cur}} \), the neighbors of this patterns (which is the set of edges \((v_i, v_j)\) such that either \( v_i \) or \( v_j \) are in \( P_{\text{cur}} \)) and the set of already visited edges. The function tries to extend the current pattern by adding one neighbor at a time and computes its commonness. If its commonness is above the threshold, it checks if the pattern is discriminating according to Definition 5 and, in such a case, adds the pattern to the current result set.

Conversely, if the pattern commonness is below the threshold, since the measure is not monotone, it is still possible to find a discriminating pattern among its super-patterns. However, the function CANRISE can evaluate if such a super-pattern can exists. Intuitively, such a function considers the best edges among the remaining ones, and checks if the pattern obtained by adding these edges to the current pattern is interesting. The pattern built in this way is not guaranteed to exist since: (i) the best edges could be not connected to the current pattern, and (ii) their strength is computed in the set of networks selected by the current pattern.

6. EXPERIMENTS

We tested our method over three real datasets coming from the Gene Expression Omnibus (GEO) public functional genomics data repository [2] that we briefly describe next.

The series GSE16134 was used in a study [22] whose aim was to investigate the disease-associated genes in periodontitis by performing a topological analysis of the differential co-expression network. It is available at the European Molecular Biology Laboratory-European Bioinformatics Institute and generated in a study [17] that investigates the association between subgingival bacterial profiles and gene expression patterns in gingival tissues of patients with periodontitis. It consists of 242 samples gingival tissues from periodontitis patients and 69 samples from healthy patients.

The series GSE25724 contains gene expression information about 7 non-diabetic and 6 type 2 diabetic people. The data have been collected for performing microarray analysis aimed at evaluating differences in the transcriptome of type 2 diabetic human islets compared to non-diabetic islet samples [8].

The series GSE65801 concerns the characterization of differentially expressed genes involved in pathways associated with gastric cancer and containing 32 gastric cancer tissues and 32 paired noncancerous tissues [12].

The three series of data are publicly available at [1].

In the following, the population consisting in unhealthy individuals is denoted as \( U \) while the population consisting in healthy individuals is denoted as \( H \).

We perform three families of experiments. First, we analyze the performances of the technique at varying the number of nodes, namely the number of genes under consideration; then, we study the sensitivity to parameters \( \tau_s \) and \( \tau_r \); finally, we discuss on discovered knowledge.

6.1 Scalability vs number of nodes

In the first set of experiments, starting from the dataset at hand, we sample [10, 50, 100, 150, 200] genes, build the associated networks and set the thresholds \( \tau_s = 0.7 \) and \( \tau_r = 0.5 \).

In Figure 1, we report the number of analyzed patterns w.r.t. the number of nodes in the network. Since the search space is related to the number of edges, in order to show the effectiveness of the pruning rules we report a curve representing the square of the number of edges. For all datasets, we can note that despite the exponential dependency of the search space on the number of edges, the number of visited patterns is much smaller than the square of the number of edges, witnessing that the search space is drastically pruned. Moreover, the figure shows that the number of patterns analyzed by the algorithm in the healthy population is much higher than the number of patterns analyzed in the unhealthy pop-
ulation. This witnesses that our approach is promising since among healthy individuals there are many regularities that are not present among unhealthy individuals.

6.2 Sensitivity to thresholds

In Figures 2 and 3 we report results of experiments aimed at analysing the sensitivity of the algorithm to parameters $\tau_r$ and $\tau_s$. In all the runs, we sample 100 genes from the dataset and we vary thresholds from 0.2 to 0.9. Since the commonness is a measure that represents how the pattern is well-supported, we imposed that the threshold $\tau_s$ is such that $\tau_s \cdot n$, where $n$ is the number of samples in the population at hand, is at least 2.

Figure 2 reports patterns over-represented among healthy individuals while Figure 3 reports patterns over-represented among unhealthy individuals.

These experiments confirm that the patterns found in the healthy population are much more than those found in the unhealthy population.

As for the strength threshold, the lower is $\tau_s$ the more are the analysed patterns and this is quite obvious since this
threshold defines the rejecting region of “representative patterns”. Conversely, as for the relevance threshold, the curves show that the number of analysed patterns is maximized for $\tau \approx 0.5$: for high values of $\tau$ such a number decreases since there are less patterns that are not considered due by chance; conversely, for low values of $\tau$, the number of analysed patterns decreases since many patterns actually due by chance are taken into account, but since their formation is likely to be random, they are present in both populations and then pruned.

### 6.3 Mined knowledge

In order to have a ground truth to consider to validate our approach, we accomplish two kinds of experiments.

First, we generate a synthetic dataset consisting in 50 healthy and 50 unhealthy individuals and 100 genes and the gene expression level is randomly generated from a normal distribution. Despite the resulting networks are quite dense (in mean we have above 4000 edges for each network) there are not interesting patterns and, indeed, the algorithm does not find any interesting edge and terminates immediately its computation.

Secondly, we consider differentially expressed genes between the two subpopulations as they provided in [2]. There, the genes are sorted by measuring their different behavior in the subpopulations. Note that this approach is deeply different from ours, since they are not able to single out discriminating relation between genes. However, we can exploit the information about interesting genes to have a validation of our approach. Thus, we extract 10 highly interesting genes and run our algorithm. Interestingly, almost all the patterns with high discriminating power (larger than 0.6) involves interesting genes. Moreover, we note that besides the information about the single gene, our patterns provide the analysts with the additional information concerning relations between such genes which are over represented among healthy individuals and under represented among unhealthy individuals.

### 7. CONCLUSIONS

We presented an approach to discover interesting graph patterns from gene expression data. In particular, in our network model there is a different labeled graph for each sample and, thus, a database of graphs for each sample set. We considered edge-labeled graphs such that each edge weight is associated to how much relevant is the co-expression between two genes, providing this way a measure of the discriminating power of each graph pattern to be detected. Preliminary results obtained on real gene expression data showed that our approach is scalable and able to single out interesting differences between healthy and unhealthy samples, also taking into account important local similarities and collaborative effects between multiple genes. In particular, it is also interesting to observe that, while the healthy case is characterized from many regularities, this is not true for the unhealthy case, in agreement with known findings on complex diseases such as cancer. Finally, our approach showed to be effective in discovering interesting genes known from the literature.

### Acknowledgments

This research has been partially supported by the PRIN project 20122F87B2 titled “Compositional Approaches for the Characterization and Mining of Omics Data” co-financed by the Italian Ministry of Education, University and Research.

### APPENDIX

#### A. Strength Computation Details

We are interested in evaluating the maxima of $f$. Since the 
logarithmic function is a continuous monotone increasing one, and since adding a constant to a function does not alter the argument of the maximum, the maxima of $f$ corresponds to the maximum of $g(x, y, \rho) = \log(f(x, y, \rho) + \log(2\pi))$. From:

\[
\log(f(x, y, \rho)) = \\
= \log\left(\frac{1}{2\pi \sqrt{1 - \rho^2}}\right) + \left(-\frac{1}{2(1 - \rho^2)}(x^2 + y^2 - 2\rho xy)\right) = \\
= -\log(2\pi) - \frac{1}{2}\log(1 - \rho^2) - \frac{1}{2} \left(\frac{x^2 + y^2 - 2\rho xy}{1 - \rho^2}\right)
\]

we have that:

\[
g(x, y, \rho) = -\frac{1}{2}\log(1 - \rho^2) - \frac{1}{2} \left(\frac{x^2 + y^2 - 2\rho xy}{1 - \rho^2}\right).
\]

Since we aim at finding the maxima of $g$ as a function of $\rho$, we consider the partial derivative of $g$ w.r.t. $\rho$:

\[
\frac{\partial g}{\partial \rho} = \frac{\rho}{1 - \rho^2} + \frac{(xy)(1 - \rho^2) - (x^2 + y^2 - 2\rho xy)\rho}{(1 - \rho^2)^2}
\]

and we look for the stationary points by calculating the values of $\rho$ where $\frac{\partial g}{\partial \rho} = 0$.

\[
\frac{\rho}{1 - \rho^2} + \frac{(xy)(1 - \rho^2) - (x^2 + y^2 - 2\rho xy)\rho}{(1 - \rho^2)^2} = 0 \implies \\
\rho^3 - \rho^2 xy + \rho(x^2 + y^2 - 1) - xy = 0 \quad (2)
\]

Therefore, we obtain a cubic equation. By setting:

\[
p = x^2 + y^2 - 1 - \frac{x^2 y^2}{3} \\
q = -xy + \frac{xy(x^2 + y^2 - 1)}{3} - \frac{2x^3 y^3}{27} \\
\Delta = \frac{q^2}{4} + \frac{p^3}{27}
\]

we obtain that the solutions of (2) depend on the sign of $\Delta$. In particular, the following wo cases are possible.
\( \Delta \geq 0 \): We have just one real solution that, as can be easily verified, corresponds to a maximum:

\[
\rho = \frac{xy}{3} + \sqrt{\frac{q}{2} + \sqrt{\Delta}} + \sqrt{\frac{q}{2} - \sqrt{\Delta}}
\]

\( \Delta < 0 \): We should compute the square root of a negative number. This task has a solution in the set of complex numbers. Let define: \( z_1 = -\frac{q}{2} + i\sqrt{-\Delta} \) and \( z_2 = -\frac{q}{2} - i\sqrt{-\Delta} \). Note that \( z_1, z_2 \in \mathbb{C} \) and \( z_2 = \overline{z_1} \). It follows that the solution of (2) can be written as:

\[
\rho = \frac{xy}{3} + \sqrt{\frac{q}{2} + \sqrt{-\Delta}} + \sqrt{\frac{q}{2} - \sqrt{-\Delta}}
\]

As there are three complex roots, there are three solutions for \( \rho \) that are solution of (2):

\[
\rho_k = \frac{xy}{3} + \sqrt{\left| z_1 \right| e^{i\frac{2\pi k}{3}}} + \sqrt{\left| z_2 \right| e^{-i\frac{2\pi k}{3}}}
\]

with \( k = 0, 1, 2 \).

It follows that there are three solutions in \( \mathbb{R} \):

\[
\rho_1 = \frac{xy}{3} \cos \left( \frac{\vartheta}{3} \right),
\rho_2 = \frac{xy}{3} \cos \left( \frac{\vartheta + 2\pi}{3} \right),
\rho_3 = \frac{xy}{3} \cos \left( \frac{\vartheta + 4\pi}{3} \right).
\]

However, not all the solutions we have found are valid stationary points for \( g \), because they are not in the function’s domain. Therefore, among the valid values of \( \rho \) (i.e. \(-1 \leq \rho \leq 1\)), we have to choose only the one that maximize \( g \).

### B. Relevance computation details

Consider Equation (2) again. Given values \( \rho_0 \) and \( y_0 \), with \( 0 < \rho_0 \leq 1 \), we aim at finding the values of \( x \) such that the value of \( \rho \) solution of Equation (2) is larger than \( \rho_0^3 \).

**Theorem 1.** Let \( \rho_0, x_0 \) and \( y_0 \) with \( 0 < \rho_0 \leq 1 \) be such that Equation (2) holds with this input. Let, also, \( x' \) and \( x'' \) be the solutions of Equation (2), solved w.r.t. \( x \), by setting \( \rho = \rho_0 \) and \( y = y_0 \). For any \( x' \leq x \leq x'' \), the value of \( \rho \) such that Equation (2) holds is greater than \( \rho_0 \).

**Proof.** First of all note that Equation (2) is a quadratic equation w.r.t. \( x \) and then it cannot admit more than two solutions. Since \( \rho_0 \) is the solution of the equation when \( y = y_0 \), Equation (2) admits for sure real solutions \( x' \) and \( x'' \).

Consider the function \( \Psi(\rho, x) \) obtained from Equation (2) by setting \( y = y_0 \). It implicitly defines a function \( \rho = \psi(x) \) and, by construction, \( \phi(x') = \phi(x'') = \rho_0 \).

In order to prove the theorem, we prove that \( \psi \) is concave for any \( x \) between \( x' \) and \( x'' \) and, then, the value of \( \rho \) is greater than \( \rho_0 \).

Consider the first derivative of \( \psi \) w.r.t. \( x \). Since \( \psi \) is implicitly defined, according to Dini’s Theorem,

\[
\frac{d\psi}{dx} = \frac{\partial \Psi/\partial x}{\partial \Psi/\rho} = \frac{-\rho^2 + 2\rho x - y_0}{3\rho^2 - 2\rho xy_0 + (x^2 + y_0^2 - 1)}
\]

In order to study the growth of the function, we have to solve the following system:

\[
\begin{cases}
-\rho^2 y_0 + 2\rho x - y_0 & \leq 0 \\
3\rho^2 - 2\rho xy_0 + (x^2 + y_0^2 - 1) & \geq 0 \\
3\rho^2 - \rho^2 xy_0 + \rho(x^2 + y_0^2 - 1) - xy_0 & = 0
\end{cases}
\]

First of all, note that the denominator of Equation (3) is always greater than 0.

Indeed, if \( 3\rho^2 - 2\rho xy_0 + (x^2 + y_0^2 - 1) < 0 \) then \( x^2 + y_0^2 - 1 = 2\rho xy_0 - 3\rho^2 - \epsilon \) for some \( \epsilon > 0 \). But this value make unsatisfiable Equation (4), since

\[
\rho^3 - \rho^2 xy_0 + \rho(2\rho xy_0 - 3\rho^2 - \epsilon) - xy_0 = (-\rho^3 - xy_0(1 - \rho^2) - \epsilon) < 0
\]

for any \( \epsilon > 0 \) and \( \rho \leq 1 \).

Thus, by considering just numerator, Equation (3) is satisfied for any

\[
x \leq \frac{(\rho^2 + 1)y_0}{2\rho}
\]

and the derivative is 0 for \( x = \frac{\rho^2 + 1}{2\rho} \).

By solving Equation (4) w.r.t. \( x \) we obtain solutions

\[
x' = \frac{1}{2\rho} \left(y_0(\rho^2 + 1) - \sqrt{y_0^2(\rho^2 - 1)^2 - 4\rho^2(\rho^2 - 1)^2} \right)
\]

\[
x'' = \frac{1}{2\rho} \left(y_0(\rho^2 + 1) + \sqrt{y_0^2(\rho^2 - 1)^2 - 4\rho^2(\rho^2 - 1)^2} \right)
\]

By coupling Equation (5) with Equations (6), we obtain that \( x' \) is smaller than \( \frac{\rho^2 + 1}{2\rho} \), \( x'' \) is larger than \( \frac{\rho^2 + 1}{2\rho} \) and the function increases before \( \frac{\rho^2 + 1}{2\rho} \) and decreases after \( \frac{\rho^2 + 1}{2\rho} \). Thus, \( x \) is a maximum and all the values of \( x \) in \( [x', x''] \) are such that \( \psi(x) > \rho_0 \), q.e.d.

### 8. REFERENCES


ABOUT THE AUTHORS:

Fabio Fassetti received the Laurea degree in computer engineering in 2004 and the PhD degree in system engineering and computer science in 2008, both from the University of Calabria, Cosenza, Italy. He has been an assistant professor of computer engineering at DIMES Dept., University of Calabria, Italy, since 2012. His research interests include data analysis, data management, knowledge representation and reasoning.

Simona E. Rombo is assistant professor at the Department of Mathematics and Computer Science, University of Palermo, Italy. Previously she was research fellow at DEIS of University of Calabria, and at ICAR-CNR (Cosenza, Italy). She has been visiting scientist at the National Institute of Heath (Bethesda, USA), Georgia Tech (Atlanta, USA), Purdue University (West Lafayette, USA). She co-authored more than fifty publications in international journals and conferences. Her main research interests involve bioinformatics, algorithms and data structures, and data mining.

Cristina Serrao received the Bachelor's Degree in computer engineering at University of Calabria in 2014 and she is currently studying Master's Degree programme in computer engineering at the same University. She has been a research fellow since 2015 and the main topics she is working on are about bioinformatics and data mining.
Multimodal Human Attention Detection for Reading from Facial Expression, Eye Gaze, and Mouse Dynamics

Jiajia Li, Grace Ngai, Hong Va Leong, and Stephen C.F. Chan
Department of Computing
The Hong Kong Polytechnic University
Hong Kong
{csjjli,csgngai,cshleong,csschan}@comp.polyu.edu.hk

ABSTRACT
Affective computing has recently become an important area in human-computer interaction research. Techniques have been developed to enable computers to understand human affects or emotions, in order to predict human intention more precisely and provide better service to users to enhance user experience. In this paper, we investigate into the detection of human attention level as a useful form of human affect, which could be influential in intelligent e-learning applications. We adopt ubiquitous hardware available in most computer systems, namely, webcam and mouse. Information from multiple input modalities is fused together for effective human attention detection. We invite human subjects to carry out experiments in reading articles when being imposed upon different kinds of distraction to induce them into different levels of attention. Machine-learning techniques are applied to identify useful features to recognize human attention level by building up user-independent models. Our results indicate performance improvement with multimodal inputs from webcam and mouse over that of a single device. We believe that our work has revealed an interesting affective computing direction with potential applications in e-learning.

CCS Concepts
• Human-centered computing → Human computer interaction

Keywords
Facial expression, eye gaze pattern, mouse dynamics, human attention level, multimodal interaction.

1. INTRODUCTION
Recent advances in miniature hardware have accelerated human-computer interaction research, in enabling the computer to interact better with human. Affective computing research [8][24] had gained tremendous momentum in recent years, demanding computers to understand human affects or emotions and to react accordingly in enhancing user experience. In order to recognize human affects, input signals reflecting human affects need to be acquired and processed. Under traditional KVM (keyboard-video-mouse) settings, input signals are mostly tied to keyboard and mouse dynamics. One can deduce some information about human affect from the keyboard [6][36] and the mouse [35][40], but the accuracy is not particularly high.

Webcam has become a de facto device thanks to the popularity of interactive social networking applications. A human can oftentimes deduce the emotion of a person sitting in front of a webcam to a certain degree of accuracy. Recent research in video processing and machine learning has demonstrated that human affects can be recognized via webcam video, noticeably via human facial features [42] and eye gaze behaviors [19]. Though there has been work on mind detection based on facial features and body gestures, research on cognition detection during reading is still limited in the aspects of feature recognition. There is also much work on reading behavior and the associated eye gaze behaviors [20][22][26]. Studies have shown that eye movement and eye behavior during reading is closely related to human comprehension and attention [26][27][28]. However, there are three main drawbacks in current state-of-the-art research works. First, many of them used intrusive devices, like the electroencephalography systems, to track the eye movement, or detect the user’s mental state as ground truth, through the use of electroencephalography (EEG) devices. Second, numerous methods studied how lexical and linguistic variables affect the eye gaze behavior during reading instead of performing a thorough analysis on the eye gaze pattern for a more ubiquitous and efficient affect detection. They need to rely on linguistic analysis of the materials being read by the human. Third, some other work designed user-dependent models for the affect detection which might not be able to accommodate unseen new users in practical applications, since it is often not practical to ask a new user to strain up the model before actually using it. We believe that reading tasks form a major category of computer usage for many users, especially for laymen and students, to warrant more systematic investigation.

In human computer interaction research, one would often exploit the expressive power resulted from multimodal interaction [23], in which the intention of a user is jointly specified by a plurality of input interaction modalities or signals representing the user. It could be effective in combining and fusing input signals acquired from the keyboard, the mouse and the webcam. In this paper, we investigate into the detection of human attention level when users are carrying out reading tasks based on a multimodal approach with ubiquitous hardware, namely, the webcam and the mouse, without relying on sophisticated devices such as head-mount devices, electrocardiogram devices or heart-beat belts for additional modalities. The webcam is capable of returning a stream of video frames, which is analyzed for eye gaze behavior.
recognition, face recognition and then temporal change in facial expression. The mouse is capturing its movement and clicking events, indirectly modeling the user activities of moving down a page for reading. For simplicity, we do not consider keyboard dynamics, since users in general do not utilize the keyboard in reading tasks.

We invite human subjects to carry out experiments in reading English articles, while recording the multimodal interaction data. Changes in human subject attention level are induced via the imposing of various levels of distraction during reading. We apply machine-learning techniques to identify useful features that assist in the determination of human attention level. Unlike in some other recent work relying on user-dependent models, we decide to build up the resilient user-independent model, which is more universal to different users, including unseen new users. Our results indicate that by combining the webcam and mouse inputs, there is a significant improvement in attention recognition over the use of a single modality alone. Our work demonstrates the feasibility of determining an interesting human affect, namely, attention level. It could find various applications in e-learning. For instance, animation and sound effects could be useful to attract teacher attention when a student starts to lose attention when learning. Change in materials presentation paradigm would be helpful, in a similar way as a teacher adapting to changes in perceived student attentiveness inside the classroom. Human physiological signals [14] could also be integrated into the framework with respect to human stress level during e-learning.

Our contributions in this paper can be summarized as: (1) we investigate into human attention level detection based on a most commonly occurring task, i.e., reading, without the use of sophisticated nor intrusive devices; (2) we adopt multimodal input processing to extract human facial features, eye gaze features and mouse dynamics; (3) we apply machine-learning techniques to build up user-independent models to recognize human attention level on reading tasks; and (4) we conduct experiments with human subjects to evaluate the accuracy of our approach. We believe that our work opens up a useful approach for interesting future user-computer interaction applications, for instance, in e-learning. The rest of this paper is organized as follows. In Section 2, we survey some related work in human affect recognition. In Section 3, we describe our recognition framework based on webcam video processing and mouse dynamics analysis, as well as the associated machine-learning techniques. In Section 4, we explain the experimental setups and the experimentation with human subjects carrying out reading tasks. We then evaluate the effectiveness and accuracy of our method in the next section. Finally, we conclude this paper briefly, with an outline of future work in Section 6.

2. RELATED WORK

Reading is a complex cognitive task that is closely related to reader attention level, comprehension ability, visual interest, oculomotor processing constrain, etc. From a more mechanical view point, reading can be considered a task where visual processing and sensorimotor control takes place in a highly structured visual environment [25], since the text page is less complex than scenes of visual objects. The eye is found to play an important role in reading, which is also proven in our experiments. Human cognition detection during reading has become an important research topic since reading is not only a remarkable human skill but also a good sample case to study the working of internal processes of the human mind and the external stimuli on the generation of complex human actions. However, human can get distracted when reading [1], for instance, by Instant Messaging [13]. It is therefore important to recognize the human attention level when formulating feedback for interactive applications to enhance user experience. Human reading cognition detection can contribute in applications such as e-learning by predicting the readers’ mental state through their external behavior and brain activity during the reading process.

The major source of inputs that can closely reflect human reading cognition rests with video streams, often captured via the webcam. Facial expression analysis based on webcam video stream has been applied to analyze cognitive states, psychological states, social behaviors, and social signals [10]. Most recent research on facial expression analysis has been focused on basic emotions, or prototypic emotions, including happiness, sadness, surprise, anger, disgust and fear [12]. These have also been extended to the recognition of fatigue [21], embarrassment [21], and pain [4]. Cognitive states, like agreeing, disagreeing, interested, thinking, concentrating, unsure, and adult attachment have been investigated [41]. Human affects can be recognized effectively from webcam video [17]. Facial expression recognition, being a powerful technique, also finds its application in understanding the student engagement in a classroom [2]. Cognitive engagement is found to have close relationship with a person’s cognitive abilities, including focused attention, memory, and creative thinking in learning [3].

Human behaviors are often better reflected by human-oriented signals. In reading tasks, the eye is the essential sensory organ involved, besides the brain. In particular, the eye is known to play an important role in reading by researchers, in addition to the obvious electrocardiogram signal, oriented from the human brain. In general, eye movements during reading can be categorized into saccades and fixations, which alternately occur during reading [26]. A saccade is a fast movement of the eye, which is usually in a direction parallel to that of the text. A fixation is the maintaining of the visual eye gaze on a single location. The purpose of a saccade is to locate a point of interest on which to focus, while processing of visual information takes place during fixations. Previous work has found that fixations tend to focus on long content words rather than short function words [16]. The frequency and length of the word can also affect the duration of the fixation on the word, with the gaze duration on longer or low frequency words being lengthier than that on shorter and high frequency words. In addition to eye movements, eye blinks have also been studied in conjunction with human cognition. Prior research has made use of the eye blinks as an indicator for fatigue detection. Divjak et al. [11] used eye dynamics and blinks to estimate human fatigue in computer use. They reported that primary eye fatigue indicators include the frequency and duration of blinks as well as the speed of eye closure. Techniques have been developed to accurately capture eye gaze behaviors from webcam videos [18] rather than relying on the use of proprietary external devices, such as Tobii [37]. It is also possible to derive human affects from eye gaze behaviors, such as stress level [19].

Despite the simplicity of the mouse in tracking movement and clicking events, it has been found to deliver interesting signals
Multimodal interfaces process two or more combined user input occurring forms of human language and behavior [23]. Human-actual feature extraction mechanisms for the three modalities, with multimedia system output, aiming to recognize naturally program. In the subsequent subsections, we will describe the location recognition and audio for command recognition.

A two-level facial feature extraction approach is adopted in our work. Frame-level and segment-level, as depicted in Figure 1. We perform feature extraction in each frame of a video clip and generate a set of frame-level facial feature vectors. We divide each video clip for an experimental subject carrying out a task into smaller units called segments. Each segment is composed of a good number of frames. Based on the frame-level facial feature vectors, segment-level feature extraction consolidates and generates a single segment-level feature vector to represent the whole segment. Before we perform frame-level feature extraction, we must first be able to recognize and track the human face in the video. Instead of performing face recognition from scratch for individual video frames, we adopt the face tracking approach. Once a face is recognized in a frame, we assume delta movement of the face in the subsequent frames. This can be achieved by computing for the facial landmarks and then their displacement across frames. Only in the event when the face loses track due to excessive movement (often due to large degree of head rotation) then the face needs to be recognized from scratch.

To perform frame-level facial feature extraction, we apply Constrained Local Models (CLM) [30] to track 66 facial landmarks from the video clips. This model is trained on the CMU Multi-PIE Face database [15], which contains over 750,000 images from 337 people. However, it fails to track some of the mouth movements, such as mouth corner depression. Thus, the Supervised Descent Method [39] is adopted to validate and optimize the 2D landmark locations. During CLM optimization, the 2D and 3D landmarks and other global and local parameters are adjusted iteratively until the face fitting regression model converges. Removing the rigid transformation from the acquired 3D shape compensates for the influence of out-of-plane rotation and produces the aligned 3D landmarks. Figure 2 indicates our usage of CLM to track the 66 facial landmarks.

We follow a standard approach to extract facial features, referred to as Action Units (AUs) [34]. We calculate the normalized distances and angles between the corresponding facial landmarks, which represent the direction and intensity of the facial movements, by extending AUs with only discrete intensity levels.

Table 1 summarizes the descriptions and measurements of the 20 facial features (f1 to f20) that we calculate from the 66 aligned 3D facial landmarks. Observing that the head orientation and position also play an important role in facial expression representation, we augment our feature list with 6 more head-oriented features (f21 to f26). The first three features measure head orientation with respect to x, y, and z axes in the webcam coordinate system. The remaining three features measure head position, with the face center position represented in the 2D image coordinate and the size of the face, revealing the distance between the face and the screen.

From our pilot study, we discover that variations in both head movement and lighting condition (e.g., heterogeneous...
After performing frame-level facial feature extraction, we extract from texture- and color-based features to geometry-based features introducing too many features to begin with. In our experiments, we select segments of length of 1 minute each.

Hope that this three-way representation of the frame statistical behavior is the contained within the segment over a moving window. The third inside the segment. The second behavior is the variation of frames contained with respect to the anchor frame. The third behavior is the variation with respect to an anchor frame. We hope that this three-way representation of the frame statistical variations suffices in providing us with a good sense of the macro-behavior of the user, while is simple enough without introducing too many features to begin with. In our experiments, we select segments of length of 1 minute each.

**Table 1: Facial features extracted from video**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Meaning</th>
<th>Formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{1,2,3,4}$</td>
<td>Inner and outer brow movement</td>
<td>Distance between eye brow corner and the corresponding eye corners (left &amp; right)</td>
</tr>
<tr>
<td>$f_{5,6}$</td>
<td>Eye brow movement</td>
<td>Distance between eye center and the corresponding brow center</td>
</tr>
<tr>
<td>$f_{7,8}$</td>
<td>Eye lid movement</td>
<td>Sum of distance between corresponding landmarks on the upper and lower lid</td>
</tr>
<tr>
<td>$f_{9}$</td>
<td>Upper lip movement</td>
<td>Distance between landmark 33 of nose bottom and landmark 51 of mouth outer contour</td>
</tr>
<tr>
<td>$f_{10,11}$</td>
<td>Lip corner puller</td>
<td>Distance between mouth corner and the corresponding eye outer center</td>
</tr>
<tr>
<td>$f_{12}$</td>
<td>Eye brow gatherer</td>
<td>Distance between inner eye brow corners</td>
</tr>
<tr>
<td>$f_{13}$</td>
<td>Lower lip depressor</td>
<td>Distance between landmark 8 of face contour and landmark 57 of mouth outer contour</td>
</tr>
<tr>
<td>$f_{14}$</td>
<td>Lip pucker</td>
<td>Perimeter of the mouth outer contour</td>
</tr>
<tr>
<td>$f_{15}$</td>
<td>Lip stretcher</td>
<td>Distance between the mouth corners</td>
</tr>
<tr>
<td>$f_{16}$</td>
<td>Lip thickness variation</td>
<td>Sum of distance between corresponding points on the outer and inner mouth contours</td>
</tr>
<tr>
<td>$f_{17}$</td>
<td>Lip tightener</td>
<td>Sum of distance between corresponding points on upper and lower mouth outer contour</td>
</tr>
<tr>
<td>$f_{18}$</td>
<td>Lip parted</td>
<td>Sum of distance between corresponding points on upper and lower mouth inner contour</td>
</tr>
<tr>
<td>$f_{19}$</td>
<td>Lip depressor</td>
<td>Angle between mouth corners and lip upper center</td>
</tr>
<tr>
<td>$f_{20}$</td>
<td>Check raiser</td>
<td>Angle between nose wing and nose center</td>
</tr>
<tr>
<td>$f_{21,22,23}$</td>
<td>Head orientation</td>
<td>Head orientation in 3D coordinate</td>
</tr>
<tr>
<td>$f_{24,25,26}$</td>
<td>Head position</td>
<td>Face center position in 2D image coordinate and face size</td>
</tr>
</tbody>
</table>

![Eye landmarks and features](image)

(a) facial and eye landmarks  (b) key eye landmark distances

**Figure 3:** Eye landmarks and features

The first set of segment-level features derived from the 26 frame-level facial features in Table 1 is calculated as the mean value of the features. For each feature $f_i$, we compute for each segment containing $S$ frames the average feature values of all $S$ frames inside the segment. For notational convenience, we denote this set of segment-level features as $f_i$ mean, where $f_i$ ($i \in [1,26]$) is the corresponding frame-level facial feature. Altogether, there are 26 features in this set.

The second set of segment-level features is computed based on moving windows of size $W$ (we select $W = 15$ based on the frame rate of 15 in our experiment). The frames in the segment are divided into units of $W$ frames each. For each feature $f_i$, the difference in feature values between the first frame in the window and the last frame in the window is computed. Then we compute the mean and standard deviation of the set of $S/W$ feature value differences for each feature over the segment. We denote this set of segment-level features as $f_i$ window mean and $f_i$ window std for the corresponding frame-level feature $f_i$. There are a total of 52 features in this set.

The third set of segment-level features is computed based on a special anchor frame. In particular, we adopt the face in the first frame of the video as the neutral face and consider changes in face in other frames with respect to this neutral face (i.e., delta face). In this third set of features, we consider the face as a whole instead of individual features. As a result, we compute one single value for the face in each frame with respect to the anchor frame (first frame) for the neutral face. We treat those 26-element feature vectors for each frame as a unit, and compute the Euclidean distance between the feature vector of frame $F_i$ and that of first frame $F_1$. This will give us $S−1$ Euclidean distances for a segment of size $S$. Finally, we compute the mean and standard deviation of those $S−1$ distances to result in only 2 global features. These two features are denoted as face mean and face std.

### 3.2 Eye Gaze Features

As illustrated in Figure 1, we extract eye gaze features from the webcam videos by eye gaze tracking and eye gaze behavior recognition. In this paper, we analyze three kinds of eye gaze behaviors for reading attention detection, including eye blinks, eye fixations and eye saccades. Before this can be done more precisely, we need to estimate the position of the pupil center of each eye, as well as extracting some other useful eye landmarks.

As presented in Section 3.1, the face CLM consists of 66 facial landmarks. Out of them, we identify 6 landmarks associated with the contour for each eye. This is depicted in Figure 3a, inclusive of 4 around the eye in red circles and 2 at the corners of the eye in green circles with red border. In order to accurately describe the eye gaze behaviors, it is crucial to properly locate the pupil center, which often cannot be detected from the appearance information of the eye region in unconstrained situations, reflected by the illumination and camera exposure) have posed significant challenges for the appearance-based features, especially with elderly people with natural wrinkles. As a result, we move away from texture- and color-based features to geometry-based features which are more resilient to variation to movement and illumination. This has significantly enhanced the robustness of our model in real-usage situations in the presence of uncontrollable environmental variations. The use of geometric facial features has effectively mitigated the noise arising from the textural and appearance channels.

After performing frame-level facial feature extraction, we extract three kinds of segment-level facial features based on the frame-level facial feature vectors reflecting different statistical behaviors. The first behavior that we are interested in is the average frame inside the segment. The second behavior is the variation of frames contained within the segment over a moving window. The third behavior is the variation with respect to an anchor frame. We hope that this three-way representation of the frame statistical variations suffices in providing us with a good sense of the macro-behavior of the user, while is simple enough without introducing too many features to begin with. In our experiments, we select segments of length of 1 minute each.
in each frame. As illustrated = < as the mean and vector can then be constructed from (3)

Upon identifying eye blinks, we need to classify the remaining eyelid closure event sequences. We first recognize eye blinks according to the value of $d_8$ in Figure 3a. As in previous studies, an eye blink is defined as eyelid closure for a duration of 50 to 500 ms [31]. Given the eye openness of each frame in a video segment, eyelid closure events can be easily detected by identifying the moments when the eye openness value of each eye goes down to 0. The sequences of eyelid closure events with duration shorter than 50 ms or longer than 500 ms are discarded as noises, which may be caused by the occasional tracking failure of the eye CLM or the turning away of the subject’s head. The remaining eyelid closure event sequences are considered as eye blinks. The duration of the eye blink is the length of corresponding eyelid closure sequence.

Upon identifying eye blinks, we need to classify the remaining eye gaze behaviors into eye fixations and saccades, namely, whether the eye gaze is focused on a word for mental processing, or moving for reading. To distinguish fixations and saccades, we analyze the horizontal and vertical movements of both eyes. For each eye, we compute the relative eye gaze position within the eye, independent on the actual coordinates of the eye in the frame. These relative horizontal and vertical eye gaze positions for each eye are computed as $d_5$ and $d_6$, in Figure 3b accordingly. From these landmark distances for each eye, we would like to establish the eye geometry, namely, the eye openness, the relative horizontal position and vertical position of the eye gaze. Eye openness is employed in the detection of the eye blinks, whereas temporal changes in the horizontal and vertical positions of the eye gaze are adopted in the detection of eye fixations and saccadic movements.

We first recognize eye blinks according to the value of $d_5 + d_6$ of each eye, which represents the eye openness as shown in Figure 3a. As in previous studies, an eye blink is defined as eyelid closure for a duration of 50 to 500 ms [31]. Given the eye openness of each frame in a video segment, eyelid closure events can be easily detected by identifying the moments when the eye openness value of each eye goes down to 0. The sequences of eyelid closure events with duration shorter than 50 ms or longer than 500 ms are discarded as noises, which may be caused by the occasional tracking failure of the eye CLM or the turning away of the subject’s head. The remaining eyelid closure event sequences are considered as eye blinks. The duration of the eye blink is the length of corresponding eyelid closure sequence.

Based on the 6 landmarks identified from the face and 9 landmarks from the eye CLMs (a total of 13 landmarks), we can compute the 6 key eye landmark distances, $d_1$ to $d_6$, in Figure 3b accordingly. From these landmark distances for each eye, we would like to establish the eye geometry, namely, the eye openness, the relative horizontal position and vertical position of the eye gaze. We first recognize eye blinks according to the value of $d_5 + d_6$ of each eye, which represents the eye openness as shown in Figure 3a. As in previous studies, an eye blink is defined as eyelid closure for a duration of 50 to 500 ms [31]. Given the eye openness of each frame in a video segment, eyelid closure events can be easily detected by identifying the moments when the eye openness value of each eye goes down to 0. The sequences of eyelid closure events with duration shorter than 50 ms or longer than 500 ms are discarded as noises, which may be caused by the occasional tracking failure of the eye CLM or the turning away of the subject’s head. The remaining eyelid closure event sequences are considered as eye blinks. The duration of the eye blink is the length of corresponding eyelid closure sequence.

Upon identifying eye blinks, we need to classify the remaining eye gaze behaviors into eye fixations and saccades, namely, whether the eye gaze is focused on a word for mental processing, or moving for reading. To distinguish fixations and saccades, we analyze the horizontal and vertical movements of both eyes. For each eye, we compute the relative eye gaze position within the eye, independent on the actual coordinates of the eye in the frame. These relative horizontal and vertical eye gaze positions for an eye are computed as $d_5$ and $d_6$, in Figure 3b accordingly. As illustrated in Figure 3a, the movements of the eyes over a temporal period can be analyzed from the eye gaze position sequence. Considering that for most human, both left and right eyes move together, we thus simplify the representation of eye gaze position by computing the mean value of the eye gaze positions of the left and right eyes. The eye gaze position sequence can then be represented as $EG = < EG_{1}, ..., EG_{k} >$, where $EG_{1}$ is the horizontal component of the eye gaze position of the $k$th item in the sequence, defined as the average of the horizontal positions of the two eyes, and $EG_{6}$ is the corresponding vertical component, as the average of the vertical positions. The movement of the eye gaze is measured as the Euclidean distance between the corresponding eye gaze points in the eye gaze sequence $EG$.

Eye fixations are defined to be periods in which the eye gaze remains stationary on a specific location. However, due to the inherent error of the eye CLM model and head movement, detecting fixations from the eye gaze signal $EG$ becomes more than simply looking for periods during which the eye gaze positions do not change. To determine the extent of noises on fixation detection, a pilot study was carried out to analyze the samples of gaze fixation on a single word. The eye gaze position sequences were calculated and the eye gaze movements between successive frames were analyzed to estimate the potential impact of noises. Let us define $mean_{move}$ and $std_{move}$ as the mean and standard deviation of the eye gaze movements detected by the eye CLM model between successive frames for the periods of eye fixation. To filter the noise exerted on the eye gaze signal, we define $\tau_{FA}$ as the fixation amplitude threshold, where

$$\tau_{FA} = mean_{move} + 3 \times std_{move}$$

Define $D_{i}$ as the eye gaze movement between successive eye gaze points $EG_{i}$ and $EG_{i+1}$, a vector $F$ can then be constructed from $EG$, where

$$F_i = \begin{cases} 1, & D_i < \tau_{FA} \\ 0, & D_i \geq \tau_{FA} \end{cases}$$

This gives us a binary vector in which elements with a value of 1 correspond to the moments when the eye gaze could be considered to be stationary. Since it has been found that fixations are rarely less than 100 ms and usually in the range of 200 to 400 ms [29], we label fixations as continuous stationary sequences that last for longer than 100 ms but shorter than 500 ms. Once the eye blinks and eye fixations have been identified, the sequences in between the fixations with duration shorter than 200 ms are considered as saccadic eye gaze movements as defined in [32]. The duration of a fixation and a saccade is the length of the corresponding eye gaze sequence.

After the three different eye gaze behaviors have been identified from the sequence of eye gaze positions, we construct the 9 statistical features that will be used to describe these three behaviors as shown in Table 2.

According to our observation of the eye gaze behaviors, the eye fixation is very indicative of the human attention level. It is notable that a reader tends to have long fixations while paying high attention to reading. This implies the reader makes efforts to process the information from the reading materials. In contrast,
short fixations happen when the reader’s attention level is low. The fixation rate is also important. Readers at low attention level usually read repeatedly until they fully understand the reading materials, which results in a high fixation rate. Besides eye fixations, eye blinks and saccades may also contribute to our research problem. Previous studies [11] have shown that eye blinks are correlated with human cognition, such as fatigue. Eye saccades can reflect the reading speed, which is closely related to the attention level.

### 3.3 Mouse Dynamics Features

Mouse dynamics have been shown to provide indicative information for affect detection in various research works [35][40]. In this paper, we attempt to relate mouse dynamics with human reading attention level, by analyzing typical mouse dynamics, including mouse click, mouse movement and mouse scrolling. Similar to facial expression recognition, we process raw mouse events to establish mouse dynamics over time. We then extract features representing mouse dynamics for each segment to align with the segment in the video clip. This enables signal fusion among the different modalities, namely, mouse signals and webcam signals.

As depicted in Figure 1, we pre-process the mouse activity log to clean extreme data values that may be due to noise. We then extract mouse patterns and then compute the actual features reflecting the mouse dynamics at the segment granularity. For instance, we compute the total distance traveled by the mouse by summing up the individual Euclidean distances traveled throughout the segment for each pair of sampled mouse coordinates. Similarly each pair of mouse coordinates indicates a mouse moving direction and the change in mouse movement direction is computed as the absolute difference in angle between the directions indicated by two consecutive pairs of mouse coordinates. Mouse scrolling features are computed based on the log of scrolling events, each of which occurs when the wheel is scrolled one discrete step. Consecutive scrolling events occurring within 1 second are considered to belong to the same scroll when the scroll step size is computed. The set of features extracted for mouse dynamics is depicted in Table 3, which can be categorized into three types: mouse click ($m_1$), mouse movement ($m_2$), and mouse scrolling ($m_{4,5,6}$), generated from the three mechanical components of the mouse (button, trackball and scroll wheel).

We notice that mouse direction is an important feature in demonstrating the “roughness” of the user. A conscious user would normally move the mouse in relatively straight lines without many changes in directions. Rapid directional changes

<table>
<thead>
<tr>
<th>Feature</th>
<th>Meaning</th>
<th>Formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_1$</td>
<td>Mouse click</td>
<td>Number of mouse clicks</td>
</tr>
<tr>
<td>$m_2$</td>
<td>Mouse distance</td>
<td>Distance traveled by the mouse in pixels  over the screen</td>
</tr>
<tr>
<td>$m_3$</td>
<td>Mouse direction</td>
<td>Amount of change in direction encountered by the mouse in angle</td>
</tr>
<tr>
<td>$m_{4,5}$</td>
<td>Mouse scroll count</td>
<td>Number of scrolls and number of changes in scroll direction (up and down)</td>
</tr>
<tr>
<td>$m_6$</td>
<td>Mouse scroll step size</td>
<td>Number of discrete steps per scroll</td>
</tr>
<tr>
<td>$m_7$</td>
<td>Mouse scroll speed</td>
<td>Average speed of mouse scrolls (step size over time period of scroll)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Attribute</th>
<th>Eye gaze</th>
<th>Attribute</th>
<th>Mouse dynamics</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{4}$ mean</td>
<td>lip pucker</td>
<td>$e_5$</td>
<td>fixation duration</td>
<td>$m_6$</td>
<td>scroll step size</td>
</tr>
<tr>
<td>$f_{5}$ window mean</td>
<td>eye brow movement</td>
<td>$e_4$</td>
<td>fixation rate</td>
<td>$m_3$</td>
<td>mouse direction</td>
</tr>
<tr>
<td>$f_{6}$ window std</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_{7}$ window std</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_{8}$ window std</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>face std</td>
<td>whole face</td>
<td>$e_1$</td>
<td>blink rate</td>
<td>$m_2$</td>
<td>mouse distance</td>
</tr>
<tr>
<td>$f_{9}$ window mean</td>
<td>eye lid movement</td>
<td>$e_2$</td>
<td>blink duration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_{10}$ window std</td>
<td>inner eye brow movement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_{11}$ window std</td>
<td>head position</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We adopt the wrapper method for feature selection which is reported to outperform filter method by considering the relationship between different features and selecting one feature subset that is best for the chosen classifier [33]. We adopt the best first searching approach for its efficiency, based on the Linear Support Vector Machine (SVM) for classification. This filtering step is very efficient in reducing the set of potential facial features from 80 down to 11. In other words, many of the original 80 features would not contribute much to the recognition task, manifested by the fact that recognition performance is not affected upon their removal. The list of potentially useful facial features is depicted in Table 4.

In Table 4, the ranking indicates the relative importance of the single feature contributing to recognition. It is simply computed as the percentage of training sets that the feature is selected for classification. Note that features in pair form are often of similar
Two are undergraduates and four are graduates, whereas other subjects’ facial expressions and mouse dynamics are better captured when using a simple regression-like algorithm in eliminating weak features. This model may not be compatible with other features in a way that putting them together may actually lower the accuracy. That is why a simple regression-like algorithm is used in our model.

4.1 Participants and Experiment Setup
We have recruited 6 subjects aged between 22 and 30, averaging 25.5. Two are undergraduates and four are graduates, whereas four are female and two are male. According to the pre-experiment survey, all subjects are skilled in using computer and capable of reading in English though their English ability varies. All are non-native speakers; the native language of two subjects is Mandarin while that of the other four subjects is Cantonese. This dictates the choice of the distracting vocal stimuli used in the experiments. Although they share the common written Chinese characters, the two dialects differ enough that speaker in one dialect without proper training or sufficient immersion would have much difficulty in understanding the other.

The experiment is carried out in a common office environment in the CHI Lab. As shown in Figure 4, the standard setup of the experiment consists of a 22-in flat LCD screen with a resolution of 1680 × 1050 pixels for displaying the articles to read, a common webcam fixed on the top of the display to record the subjects’ face and upper body, and a common wired mouse. All the devices are non-intrusive to the subjects. The light in the room is adjusted to be suitable for reading and is maintained stable throughout the experiment. The subjects are seated about 60 cm away in front of the display.

Data collection programs run on the computer displaying the article for reading. Both the content shown on the screen and the webcam vision are recorded by a free-trial version of the software Camtasia, capturing the two video streams at a frame rate of 15 per second onto the hard disk. We develop a C++ program to capture and log mouse events to determine mouse dynamics, including mouse click, mouse scrolling and mouse movement, together with their timestamps. Mouse click and scroll events are logged when they occur, and mouse coordinate is sampled at a rate of approximately 15 per second for mouse movement. The program is run concurrently and the timestamped information is stored in the hard disk for temporal alignment.

4.2 Experiment Design
In our experiment, each subject is required to read three different English articles chosen from TOFEL (Test of English as a Foreign Language) reading comprehension materials. We decide to select articles from TOFEL because the topic, length and difficulty of the articles are proper for our non-native speaker subjects in this reading experiment. The time spent on reading is not constrained. In this paper, a reading session refers to the particular experiment in which one subject reads one TOEFL article. To make sure that the subjects really read the articles with reasonable amount of efforts instead of just killing time, they are required to write a short summary of at least 50 words after finishing reading the article in each session.

In the first set of sessions, the subjects read in a quiet environment without anything to distract them. To induce different levels of attention on the subjects, we choose two kinds of vocal stimuli to distract the subjects on purpose during reading in the second and third set of sessions. One of the vocal stimuli is heavy metal music which carries a “high information-load” and supposedly to be able to impair performance significantly in reading comprehension task [5]. The other vocal stimuli is sound recording of famous funny talk shows that the reader would very likely be interested in. Considering the different native languages of our subjects, we choose Mandarin and Cantonese talk shows for the subjects based on their native language. By doing this we make sure that all the subjects can understand the contents of the talk shows easily even in the background, so as to distract them.
watching video clips \[38\]. A user-independent model can be applied on new users, but the amount of data needed for training the classifier will be much larger. User-independent models can be built for each individual subject and thus remove those occasional corrupted video data. The amount of such bad data only contributes to less than 10% of the total data. Finally, we are able to collect data of a length of 147 minutes for all the six subjects (about 25 minutes per subject).

We next establish the ground truth and baseline from the dataset for evaluation purpose. According to the attention level labeled by our subjects, 35.4% of the data is labeled as “high”, 34.7% of the data is labeled as “medium” and 29.9% of the data is labeled as “low”. This is a set of very well-mixed data, since the three classes are roughly equally represented without much skewness. The baseline of the dataset is 35.4%, since the bottom line for random guessing in classification is to output the label of the largest class for a “best” result. This baseline of a dataset is widely used to evaluate the classification performance of an algorithm. In this paper, we build up user-independent models for attention detection based on this dataset.

### 4.3 The Dataset

We have to perform pre-processing to the video clips since there are occasional instances with subjects showing only a partial face, caused by inappropriate sitting position of the subject. As our facial expression model depends on the key landmarks throughout the face, a partial face without the mouth would not be useful. We thus remove those occasional corrupted video data. The amount of such bad data only contributes to less than 10% of the total data. Finally, we are able to collect data of a length of 147 minutes for all the six subjects (about 25 minutes per subject).

We take the video data and preprocess it. Additionally, we only use the video data by leaving out a different subject and the average performance is reported. The confusion matrix normalized by the ground truth is shown in Table 5. We can see that features describing the change of frame-level feature vectors are mostly chosen. This indicates that the magnitude of the change of facial expression of specific areas on the face varies with the level of attention of the subjects. Within all the selected features, the change of eye brow position and eye brow movement are particularly important when compared with other features.

Since we are building user-independent models, the gold standard in evaluating the effectiveness is the leave-one-subject-out cross-validation test. From the set of \(n\) subjects, we train the user-independent model with dataset from \(n-1\) subjects and test the model on the left-out subject. We repeat the experiment \(n\) times by leaving out a different subject and the average performance is reported. The confusion matrix normalized by the ground truth and the performance matrix for classification are shown in Table 6 and Table 7.

### 5. EVALUATION

In this section, we evaluate our multimodal attention detection approach by building user-independent models based on the combined dataset of all subjects. In classification research, a user-independent model is usually not as accurate as a user-dependent model, but is more applicable in practice. The former can be built easily but the latter has to be built for each individual subject and the amount of data needed for training the classifier will be much larger. User-independent models can be applied on new users, but user-dependent models cannot. We build up user-independent models and evaluate our approach in this section.

We compare the classification performance of our multimodal approach with the performance produced by using only a single modality. The results illustrate that the multimodal models perform better than the single modality ones, achieving higher correct classification rates (CCR) and F-measures.

#### 5.1 Attention Detection with Facial Features

Our first evaluation is concentrated on the use of facial features extracted from the webcam video to recognize attention level in reading tasks. There are a total of 80 extracted facial features across three categories, trimmed down to 11 via the wrapper approach upon adopting the Linear Support Vector Machine (SVM) with 10-fold cross-validation to classify the dataset.

From the set of 11 potential facial features, we attempt different subset combinations and select 7 producing a best performance, as shown in Table 5. We can see that features describing the change of frame-level feature vectors are mostly chosen. This indicates that the magnitude of the change of facial expression of specific areas on the face varies with the level of attention of the subjects. Within all the selected features, the change of eye brow position and eye brow movement are particularly important when compared with other features.

From Table 6 and 7, it can be observed that the average CCR for the three classes is 68.7%, and this is significantly higher than the baseline of 35.4% with an improvement of 33.3% (doubling the accuracy). It can also be seen that most of the errors come from misclassifying as the neighboring attention level class, i.e., low \(\leftrightarrow\) medium and medium \(\leftrightarrow\) high. Only very few errors are due to misclassification of extreme classes between low \(\leftrightarrow\) high. Similarly, we are able to achieve a high precision as well as a high recall, without having to sacrifice one metrics for the other. The resultant F-measure is also as high as 0.7, close to the CCR.
5.2 Attention Detection with Eye Gaze Features

We believe that the 9 eye gaze features will not contribute equally to the attention level classification. To explore the most indicative set of eye gaze features, we compare the classification performance with different combination of eye gaze features and find out 5 useful eye gaze features, as shown in Table 5. Those 5 eye gaze features are $e_1$, the rate of eye blinks, $e_2$, the average blink duration, $e_4$, the rate of eye fixations, $e_5$, the average fixation duration, and $e_6$, the average saccade duration. It is worth noticing that features representing all three kinds of eye gaze behaviors analyzed in this paper are selected in the subset. It indicates that there is a strong correlation between the eye gaze behaviors and the level of attention in our reading task. Moreover, the top two ranked features are both eye fixation features, which validates our findings that eye fixation is critical to the attention level detection in our work. According to Table 5, none of the eye gaze features representing the standard deviation of the eye gaze behaviors ($e_3$, $e_4$ and $e_5$) is selected. It perhaps implies that the eye gaze behavior patterns are quite stable with a certain attention level. We build a user-independent model based on eye gaze features alone and perform the leave-one-subject-out cross-validation test. The confusion matrix normalized by the ground truth and the performance matrix for classification are shown in Table 8 and Table 9.

As shown in Table 8 and 9, the averageCCR for the three classes is 58.5%, which is higher than the baseline by 23.1% with only 5 features. Similar with the facial feature model, the CCR of the low class is better than that of the medium class, while the high class is still the one with biggest misclassifying errors. Although the errors still mainly come from misclassifying between low ↔ medium and medium ↔ high as in the facial feature model, we note that the error to misclassify high as low becomes bigger than the facial feature model. It means the eye gaze behaviors analyzed in this paper do not correspond that well with the level of attention as with facial expressions. This may sound intuitive, since the facial expression carries inherently richer information than the eye gaze alone. Nevertheless, the eye gaze features still contribute a lot to the attention level classification, despite its relatively small amount of features and landmarks required. Finally, the average recall and precision for the three classes are 0.57 and 0.62 respectively, whereas the average F-measure is 0.57, consistent with the CCR and somewhat lower than those performance based on facial features.

5.3 Attention Detection with Mouse Dynamics

There are only 7 mouse dynamics features but not all of them contribute well to the classification process. We therefore explore different subsets of feature combinations for mouse dynamics and we land on 3 useful mouse features for classification as shown in Table 5. Those useful features are $m_4$, the amount of scrolling steps, $m_3$ and $m_2$, the amount of changes in mouse direction and total distance that the mouse travels. In our experiment, we observe that the mouse click events are not indicative at all. This is because most subjects only use the mouse scrolling button to navigate up and down the article, instead of clicking on the scrollbar in the application window in this reading task. The distance traveled and direction changed for the mouse come up as important features contributing to the classification of the attention level. The mouse click events would be more useful when writing tasks are studied, so would keyboard dynamics be. In any case, the selected features demonstrate that the mouse trajectory is indicative for attention level classification. The normalized confusion matrix on classification and its accuracy based on mouse dynamic features is depicted in Table 10 and Table 11.

According to Table 10 and 11, the average CCR for the three classes is around 44.9%, which is not as good as the performance of the facial feature model and the eye gaze feature model. When compared with the baseline of 35.4%, there is still an improvement of 9.5%, even with as few as 3 mouse features. Although the improvement is not as impressive when compared with those of facial features, the result is already acceptable with just 3 features. We believe that the lack of useful information about the mouse dynamics during the reading task drags the classification performance to a certain extent. It can also be observed that there are more classification errors across extreme classes, i.e., low ↔ high. This is perhaps due to the fact that mouse dynamics do not correspond that well with the attention level as with facial features and eye gaze features. Nevertheless, the recall and precision metrics and the F-measures for the three classes remain stable at about 0.45, similar to the CCR.

5.4 Attention Detection with Multimodalities

We have already observed good recognition with the unimodal models based on facial features and acceptable recognition based on eye gaze behaviors and mouse dynamics in our study. We now adopt the multimodal model by combining the features of all the modalities. There are a total of 15 features in this multimodal recognition study as shown in Table 5. We build user-independent
models based on SVM and apply 10-fold cross-validation in the evaluation. As before, we employ the challenging leave-one-subject-out cross-validation experiment over the n subjects. The confusion matrix normalized by the ground truth and the performance matrix for classification are shown in Table 12 and Table 13. From the two tables, the average CCR for the three classes is found to be 75.5%, an improvement of 40.1% over the baseline, with the accuracy of one class going up to 81%. Although the classification performance based on mouse dynamics is much lower than that one based on facial features or eye gaze features, the overall performance has been improved compared with individual performance, when the three modalities are combined. The classification errors across neighboring classes and especially the extreme classes of low ↔ high have all been reduced when compared with the use of features of single modality. It is also worth noticing that the performance of the medium class improves dramatically compared with the eye gaze feature model and the mouse dynamics model. When we look at the recall and precision metrics, they show similar pattern as that of CCR with a comparable F-measure. In summary, we believe that our selected features of different modalities contribute to the attention level detection during reading in a synergic way.

5.5 Contributions by Individual Modalities

We can attain different performance based on features generated from each individual input modality. From Section 5.1 to 5.3, it is easily seen that facial features produce the best performance, followed by eye gaze features and finally mouse features. However, in terms of computational cost, the reverse is true. This is the rationale behind the choice of a proper multimodality feature set to yield a good enough recognition rate. In this section, we proceed to analyze more deeply the individual contribution by each modality and see which combinations would produce a better integrative performance.

We conduct three more experiments based on (a) combined facial and eye gaze features, (b) combined facial and mouse features, and (c) combined eye gaze and mouse features. From there, we would be able to identify the contribution by individual modality more precisely. The corresponding confusion matrices for the three combinations are summarized in Table 14. It can be seen that they exhibit intermediate performance with respect to those for single component modalities and the one for the full set of modalities, as compared with those in the previous tables. The precision/recall metrics show a similar pattern as in the previous experiments and are thus omitted.

For comparison, we report the CCR for these combinations, alongside those of the individual feature sets. We also compute the improvement in CCR performance for each combination. This improvement indirectly measures the “synergic” effect between the two feature sets. It is conceivable that a higher synergic effect is more preferred. The results are depicted in Table 15.

We can observe from Table 15 that facial features integrate well with mouse features to produce a best improvement of 16% in terms of CCR performance, whereas the other two combinations only produce about 10% improvement. This observation yields a slightly different conclusion based on absolute performance alone, which suggests that the model based on facial features combined with eye gaze features performs the best at 73.5% against 72.8% for facial features combined with mouse features. Nevertheless, this higher performance is attained at the expense of adopting the higher cost eye gaze feature set than the lower cost mouse feature set.

Let us make a simplifying assumption that all feature sets are somewhat synergic to one another, in order for us to take a glance on the contributions by the individual modalities. In other words, we assume that the models would not have negative impact on one another when combined. The synergic effect is much higher than the interference effect. We can then attempt to break down for the individual contributions based on a simple additive model as shown in Figure 5. This provides us with a glance on the individual contribution to the overall performance. The more performance that can be “explained” by the overlapping part of two models, the more “similar” are the two sets of features and the higher possibility that the two models are making similar classification. As a result, there would be less additional improvement incurred in the multimodal model. Finally, it can be seen that any of the three models alone would produce an accuracy of close to 40%, which accounts for more than half of the attainable performance for the three models. Actually, this already represents the majority of the performance for the mouse feature model. This is a pretty high degree of “similarity” among the three individual models. Also, the “similarity” between facial feature model and eye gaze feature model is relatively high and this is understandable, as both come from the same video captured by the webcam.

<table>
<thead>
<tr>
<th>Facial + eye gaze</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR</td>
<td>0.79</td>
<td>0.15</td>
<td>0.06</td>
</tr>
<tr>
<td>CCR</td>
<td>0.25</td>
<td>0.73</td>
<td>0.02</td>
</tr>
<tr>
<td>CCR</td>
<td>0.11</td>
<td>0.20</td>
<td>0.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Facial + mouse</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR</td>
<td>0.83</td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td>CCR</td>
<td>0.22</td>
<td>0.69</td>
<td>0.10</td>
</tr>
<tr>
<td>CCR</td>
<td>0.16</td>
<td>0.18</td>
<td>0.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eye gaze + mouse</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR</td>
<td>0.83</td>
<td>0.15</td>
<td>0.02</td>
</tr>
<tr>
<td>CCR</td>
<td>0.37</td>
<td>0.49</td>
<td>0.14</td>
</tr>
<tr>
<td>CCR</td>
<td>0.18</td>
<td>0.27</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 14: Normalized confusion matrix for multimodal models

<table>
<thead>
<tr>
<th>A+B</th>
<th>CCR</th>
<th>CCR</th>
<th>ΔA</th>
<th>ΔB</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>facial+eye gaze</td>
<td>73.5%</td>
<td>68.7%</td>
<td>4.8%</td>
<td>15.0%</td>
<td>9.9%</td>
</tr>
<tr>
<td>facial+mouse</td>
<td>72.8%</td>
<td>68.7%</td>
<td>4.1%</td>
<td>27.9%</td>
<td>16.0%</td>
</tr>
<tr>
<td>eye gaze+mouse</td>
<td>62.6%</td>
<td>58.5%</td>
<td>4.1%</td>
<td>17.7%</td>
<td>10.9%</td>
</tr>
</tbody>
</table>

Figure 5: Improvement breakdown against models

Table 15: CCR improvement for individual modalities
5.6 Performance for Existing Users

So far in all our evaluations, we assume the setting of leave-one-subject-out for recognition performance to cater for unseen new users. It is also common in reality that the model is used by an existing user. One would expect that the accuracy will be higher. In our next experiment, we keep all subjects in the 10-fold cross-validation and compare the performance with the leave-one-subject-out setting as presented in Table 16. We observe a bit of improvement in terms of CCR. On the other hand, this small improvement also demonstrates that our approach is very robust, in delivering good performance even for unseen new users based on training data from just a small number of subjects \( n-1 = 5 \). Table 17 and Table 18 provide more information of the recognition performance of the multimodal model with 10-fold cross-validation. Compared with Table 12 and Table 13, we observe that there are improvements of the CCR for each class, and the misclassifying errors between neighboring classes and extreme classes are further decreased. Thus, our research represents a good initial attempt to attention detection based on ubiquitous devices and a small number of features extracted from webcam videos and mouse activities.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose to recognize human attention level via the use of ubiquitous equipment loaded with most computers, namely, the mouse and the webcam. We extract facial features and eye gaze features from the videos captured by the webcam, as well as mouse dynamics due to mouse usage. We adopt machine-learning techniques on modeling the data captured and build user-independent models capable of recognizing the attention level for unseen new users. We conduct our experiments via the reading tasks, with which the subjects are induced to different levels of attention, through the use of different vocal stimuli to distract them. Our results based solely on the webcam (i.e., facial features and eye gaze features) indicate good performance, and those solely on the simplistic mouse still achieve improvement over the baseline. Finally, we demonstrate that combining the three sets of features together is giving us the best performance, whereas only 15 important features need to be utilized.

In the future, we would like to study the improvement by means of user-dependent models, upon collecting a larger dataset. Furthermore, we would like to build up a number of good user-independent models based on an initial classification on the user category. We could then start from those more specific user-independent models to move towards more accurate models tailored made for individual users. We would also like to expand our scope of investigation to include writing tasks, making use of keyboard dynamics, so as to integrate the three ubiquitous modalities, namely, webcam video, mouse dynamics and keyboard dynamics to cover the major categories of user activities, namely, reading and writing tasks, that take up a huge chunk of computer usage.

7. ACKNOWLEDGMENTS

The authors would like to thank the subjects who participated in the experiments. This research is supported in part by the Hong Kong Research Grant Council and the Hong Kong Polytechnic University under Grant numbers PolyU 5222/13E and PolyU 152126/14E.

8. REFERENCES


**Table 16: CCR improvement for existing users**

<table>
<thead>
<tr>
<th>Model</th>
<th>Facial</th>
<th>Eye gaze</th>
<th>Mouse</th>
<th>Multimodal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leave-one-subject-out</td>
<td>68.7%</td>
<td>58.5%</td>
<td>44.9%</td>
<td>75.5%</td>
</tr>
<tr>
<td>All-subjects-included</td>
<td>70.1%</td>
<td>61.2%</td>
<td>45.6%</td>
<td>78.9%</td>
</tr>
</tbody>
</table>

**Table 17: Normalized confusion matrix for existing users**

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.83</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td>Medium</td>
<td>0.10</td>
<td>0.82</td>
<td>0.08</td>
</tr>
<tr>
<td>High</td>
<td>0.09</td>
<td>0.20</td>
<td>0.70</td>
</tr>
</tbody>
</table>

**Table 18: Classification performance for existing users**

<table>
<thead>
<tr>
<th>Attention level</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Medium</td>
<td>0.72</td>
<td>0.82</td>
<td>0.77</td>
</tr>
<tr>
<td>High</td>
<td>0.84</td>
<td>0.70</td>
<td>0.77</td>
</tr>
</tbody>
</table>


[37] Weigle, C. and Banks, D.C. Analysis of eye-tracking experiments performed on a Tobii T60. Presented at the Conference on Visualization and Data Analysis, (2008), San José, California, USA.


ABOUT THE AUTHORS:

Jiajia Li received the BEng and MEng degrees in Automation Science from Qilu University of Technology and Beihang University in 2009 and 2012, respectively. Currently, she is working toward the PhD degree in the Department of Computing at the Hong Kong Polytechnic University, Hong Kong SAR, China. Her research interests include affective computing, human computer interaction, and cross-modal art generation.

Grace Ngai received her Ph.D. degree from Johns Hopkins University in Computer Science in 2001. She worked for Wenwen Technologies, a natural language and speech firm in Hong Kong, and joined the Hong Kong Polytechnic University in 2002. She is currently an associate professor at the Department of Computing. Her research interests are in affective computing, human computer interaction, wearable computing, and education.

Hong Va Leong received his PhD degree from the University of California at Santa Barbara in 1994 and joined the Hong Kong Polytechnic University. His research interests lie in distributed systems, distributed databases, and mobile computing. He has served on the program committee and organizing committee of numerous international conferences as well as chairing some of them. He is a member of the ACM, IEEE Computer Society and IEEE Communications Society.

Stephen C.F. Chan received his Ph.D. degree in Electrical Engineering from the University of Rochester in 1987. He had worked for the National Research Council of Canada, and was the Canadian representative for the ISO-10303 STEP standard for the exchange of industrial product data. He is currently an associate professor in the Department of Computing at the Hong Kong Polytechnic University. His research interests are data and text mining, human-computer interaction and service-learning.