A HYBRID RECOMMENDATION SYSTEM For NEWS IN A MOBILE ENVIRONMENT
INTRODUCTION AND MOTIVATION
The consumption of printed morning papers has decreased substantially over the last 20 years.

Alternative sources of news and information are becoming available to a potential vast audience via an increasing number of wireless devices.

The traditional news industry is changing and transforming itself into a digital and online news industry.
Overall, more people are accessing news through a greater number of devices than ever before.

The computer remains the most important device for online news, but for many, this is now supplemented by heavy usage of smartphones and tablets.

In many countries, publishers now report that the majority of traffic comes from new mobile devices.
News Recommendation System

Multiples on-line news portals confront users with the choice between numerous items inducing an information overload.

As the volume of news grows, so does the need for tools which act as filters, deliver only information that can be considered relevant to the reader.

In the face of vast amount of information, recommendation systems stand out as a possible solution to help readers on the selection of content.
Pglocal project

Pglocal: aggregating news from various sources

• Pglocal is a new concept of editorial content aggregation
  • Joins in the same platform a network of newspapers, universities and users from various Portuguese-speaking countries (Portugal, Brazil, Mozambique)

• Main goal
  • Build a community of readers who have in common the language
  • Create or consolidate a paywall for collecting digital content and subscriptions
  • Explore the information business in the international Portuguese-speaking market

• Multiple devices and operative systems (computers, android and iOS mobile devices)

• Crowdsourcing component
  • Users can write news and upload them to the system as well as images and videos captured through the smartphone
PGLOBAL
RECOMMENDATION SYSTEM
Hybrid recommendation systems combine two or more recommendation methodologies.

- **Content-Based**
  - Recommends news that are similar to the ones an user has shown interest in the past.
  - Considers the change of user interests over time.

- **Georeferenced**
  - Recommends news having the same geolocation of the mobile device.
System architecture

Pglocal News Provider
Aggregating and providing news articles from multiple Portuguese language newspaper sources

Device Logs Server
Stores device logs

Log file content:
- device ID
- read news
- shared news
- liked news
- country code

Mobile device
Runs the news App

Recommendation Core
Periodically generated, in offline mode, device recommendations. Produced recommendations are sent to the Pglocal repository in order to be made available to a mobile device
News Processing

PHASE 1
NEWS PRE-PROCESSING
Extract main topics and key information

PHASE 2
CLUSTERING
Group articles according to their content

PHASE 3
NEWS SIMILARITY
Compute the similarity between news
Astronomers have confirmed the existence of planets that have had their atmospheres stripped away by their host stars.

Planets with atmospheres that orbit too close to their host stars are bombarded by a torrent of high-energy radiation.

The gaseous outer layers of these worlds are then stripped away, according to the international team of scientists.

The researchers used data from Nasa's Kepler space telescope, which was launched to identify and study exoplanets, which circle stars other than our own.

They focused on a category of planet called "super Earths", which are roughly 2-10 times more massive than our own planet.

The astronomers used a technique called asteroseismology, which probes the internal structure of stars, to confirm the idea.

The results of the study have important implications for understanding how stellar systems, like our own Solar System, and their planets, evolved over time - including the role played by the host star.
NEWS PREPROCESSING
PHASE 1
News Preprocessing

**STEP 1**
Split the title and subtitle into **tokens**

**TOKENIZATION**
Breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens

**Planets stripped bare by host stars**
Astronomers have confirmed the existence of planets that have had their atmospheres stripped away by their host stars.

**STEP 2**
All common non informative words (stop words) are removed

**STOP WORDS REMOVAL**
Stop words usually refer to the most common words in a language
Ex: “the”, “an”, “of”, ...

**Planets stripped bare by host stars**
Astronomers have confirmed the existence of planets that have had their atmospheres stripped away by their host stars.

**STEP 3**
Reducing inflected words to their stem (or root) word

**STEMMING**
Reducing inflected (or sometimes derived) words to their word stem, base or root form

**Planets stripped bare host stars**
Astronomers confirmed existence planets atmospheres stripped away host stars.
Planets with atmospheres that orbit too close to their host stars are bombarded by a torrent of high-energy radiation. The gaseous outer layers of these worlds are then stripped away, according to the international team of scientists. The researchers used data from Nasa’s Kepler space telescope, which was launched to identify and study exoplanets, which circle stars other than our own. They focused on a category of planet called “super-Earths”, which are roughly 2-10 times more massive than our own planet. The astronomers used a technique called asteroseismology, which probes the internal structure of stars, to confirm the idea. The results of the study have important implications for understanding how stellar systems, like our own Solar System, and their planets evolved over time—including the role played by the host star.
Astronomers have confirmed the existence of planets that have had their atmospheres stripped away by their host stars. 

Planets with atmospheres that orbit too close to their host stars are bombarded by a torrent of high-energy radiation. 

The researchers used data from NASA's Kepler space telescope, which was launched to identify and study exoplanets, which circle stars other than our own. 

They focused on a category of planet called "super Earths", which are roughly 2-10 times more massive than our own. 

The astronomers used a technique called asteroseismology, which probes the internal structure of stars, to confirm the idea. 

The results of the study have important implications for understanding how stellar systems, like our own Solar System, and their planets, evolved over time - including the role played by the host star.
Clustering

- Grouping news in such a way that news in the same group (cluster) are more similar to each other than to those in other groups.

- Each cluster has assigned a human readable label.

- The same news can belong to different clusters.

- Supported by the Carrot2 Framework:
  - Organize small collections of documents into thematic categories.
NEWS SIMILARITY
PHASE 3
News Similarity

- Computing news similarities between items belonging to the same cluster
  - Lower the time needed for similarity estimation
  - Avoids calculating the similarity between items that do not have any content in common

---

**Planets stripped bare by host stars**

**Astronomers** have confirmed the existence of **planets** that have had their atmospheres stripped away by their host stars.

**Planets** with atmospheres that orbit too close to their host stars are bombarded by a torrent of high-energy radiation…

---

**Evidence suggests huge ninth **planet** exists past Pluto at solar system's edge**

**Astronomers** investigating the odd alignment of rocks beyond Pluto have concluded that an undetected icy **planet** four times the size of Earth must exist.

As science often does, it began with a “huh?” Some distant objects far beyond Pluto were behaving very oddly. The orbits of a handful of space rocks had…
News Similarity

- Similarity between news inside each cluster is estimated using the items’ weight terms vectors
  - Vector similarity estimated using Cosine similarity

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

### ITEMS’ WEIGHT TERMS VECTORS AND SIMILARITY - EXAMPLE

<table>
<thead>
<tr>
<th></th>
<th>planet</th>
<th>...</th>
<th>star</th>
<th>...</th>
<th>pluron</th>
<th>...</th>
<th>strip</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEWS 1</td>
<td>1.00</td>
<td>...</td>
<td>1.00</td>
<td>...</td>
<td>0.00</td>
<td>...</td>
<td>0.33</td>
</tr>
<tr>
<td>NEWS 2</td>
<td>0.65</td>
<td>...</td>
<td>0.00</td>
<td>...</td>
<td>0.80</td>
<td>...</td>
<td>0.00</td>
</tr>
<tr>
<td>NEWS 3</td>
<td>0.40</td>
<td>...</td>
<td>0.16</td>
<td>...</td>
<td>0.00</td>
<td>...</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>NEWS 1</th>
<th>NEWS 2</th>
<th>SIMILARITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEWS 1</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEWS 1</td>
<td>0.23</td>
<td>NEWS 3</td>
<td></td>
</tr>
<tr>
<td>NEWS 2</td>
<td>0.42</td>
<td>NEWS 3</td>
<td></td>
</tr>
</tbody>
</table>
Device Profile – Information Collecting

• Constructed by collecting information implicitly

• Implicit information results from three types of user’s interaction with the mobile application:
  • Read – Registered when a user keeps in the same news screen for a pre-defined period of time.
    • For example, more than 10 seconds.
  • Like – This action corresponds to a Facebook like that a user gives to a news article
  • Share – This action corresponds to sharing a news article on an available social network, namely, Facebook, Twitter and Google+
Device Profile - Approaches

- SHORT-TERM
- LONG-TERM
- GEOREFERENCED
Device Profile – Short-term

- Short-term profile tries to capture recent or seasonal interests
  - For example, very popular news

- Relevant in a mobility scenario
  - Users will probably modify their reading habits during short stays in other locations

- Based on the set of $N$ recent news with which the user has interacted (read, like or share)

### DEVICE SHORT-TERM PROFILE - EXAMPLE

<table>
<thead>
<tr>
<th>NEWID</th>
<th>DATE</th>
<th>ACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2016-01-10 12:20:22</td>
<td>READ</td>
</tr>
<tr>
<td>2</td>
<td>2016-01-10 09:24:14</td>
<td>READ, SHARED</td>
</tr>
<tr>
<td>3</td>
<td>2016-01-09 14:12:32</td>
<td>READ, LIKED</td>
</tr>
<tr>
<td>4</td>
<td>2016-01-07 12:34:21</td>
<td>READ</td>
</tr>
<tr>
<td>5</td>
<td>2016-01-05 18:32:22</td>
<td>READ</td>
</tr>
</tbody>
</table>
Device Profile – Long-term

• Generally, over a long period of time this profile will remain stable and not subject to major changes

• Created based on the news that a user reads over time

• System collects the news tags and the profile is created by considering the number of times that a specific tag has been read by a user
  • If a user reads 8 news containing the tag “football” that tag will have a count of 8

<table>
<thead>
<tr>
<th>TAG NAME</th>
<th>OCCURRENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>sport</td>
<td>12</td>
</tr>
<tr>
<td>football</td>
<td>8</td>
</tr>
<tr>
<td>economy</td>
<td>7</td>
</tr>
<tr>
<td>science</td>
<td>5</td>
</tr>
<tr>
<td>android</td>
<td>3</td>
</tr>
</tbody>
</table>
Device Profile – Georeferenced

• Consider the country where the device is located
  • Performed by monitoring device IP address

• Geolocation information is coded using ISO 3166-1 Alpha-2 country code

DEVICE GEOLOCATION - EXAMPLE

AUSTRALIA
ISO 3166-1 Alpha-2: AU
HYBRID RECOMMENDATION SYSTEM
Hybrid Recommendation System

- Implemented hybrid recommendation algorithm combines **content-based** (short-term, long-term) and **georeferenced** recommendations approaches.

- **Long-term**: Match of keywords/terms which are present simultaneously in the current news content and in the user keyword profile.

- **Short-term**: Recommending news with similar topics to those the users have read recently.

- **Georeferenced**: Recommending news related the country where user is located.
Hybrid Recommendation System

Short-term recommendations

- Recommending news with similar topics to those the user has read, shared or “Liked” recently
Hybrid Recommendation System

Long-term recommendations

- Recommendation algorithm matches the tags that have been cumulatively added to the profiles with the tags of existing news

<table>
<thead>
<tr>
<th>RECOMMENDATIONS</th>
<th>NEWS TAGS</th>
<th>LONG-TERM PROFILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1º</td>
<td>Politics, USA, President</td>
<td>Politics (5)</td>
</tr>
<tr>
<td>2º</td>
<td>Sport, Football</td>
<td>Football (4)</td>
</tr>
<tr>
<td>3º</td>
<td>Wall Street, Dollar, Economy</td>
<td>Economy (3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cinema(1)</td>
</tr>
</tbody>
</table>
Hybrid Recommendation System

Georeferenced recommendations

- Recommending news related to the country where the device is located
  - Recommendations that can deliver interest topics to the user, in any place, at any time, for example, while traveling on a business trip or on holidays
Hybrid Recommendation System

Final recommendation list

• Created by combining all the previous described approaches
  • Pre-defined fraction for each
    • SHORT-TERM – 60%
    • LONG-TERM – 30%
    • GEOREFERENCED – 10%

• Priority in the list is assigned to each of the recommendation methodologies
  • 1º - SHORT-TERM
  • 2º - LONG-TERM
  • 3º - GEOREFERENCED
EVALUATION AND RESULTS
Experimental Evaluation

• In order to evaluate the recommendation system, a user study was conducted
  • A set of volunteer subjects were recruited

• The user evaluation group consisted of about 35 participants

• Participants with different reading profiles:
  • seasonal readers - read only a very small number of news
  • average readers - having an intermediate amount of news accesses
  • intensive readers – having a high reading historical profile

• Participants were asked to perform 2 experiments
Experimental evaluation

- Presenting to the user three lists of news corresponding to:
  1. the output of the recommendation system
  2. the most popular news and
  3. randomly selected news.

- Each list was composed of 5 newspaper news, horizontally aligned

- The ranking position of each list was randomly generated
  - Dismiss possible ranking influence

- For this task the user was requested to select the list of news he thinks that best matches his preferences.
Results – Experiment 1

- List resulting from the recommendation algorithm was most often chosen
- Only a small difference is observed for most popular and random lists

Results are consistent independently of the position of the recommendation list in the experiment
Results – Experiment 1

- Good performance of the recommendation algorithm when considering only loyal readers

- Results are consistent independently of the position of the recommendation list in the experiment
Experimental Evaluation

- 12 news presented to the user
  - Resulting from the 3 recommendations approaches (Geo-located, Short-term, Long-term)
  - Each approach generated the same number of recommendations

- The user was asked to rate, 1 to 5, each of 12 news
Results – Experiment 2

• The short-term recommendation algorithm obtained the highest preferences (rating average)
  • Users’ higher interest for news similar to those ones he read, share, like recently

• Worst performance for the georeferenced algorithm
  • May be due to the use of only generic information about the country where the user is located

• Results also contributed to define the hierarchy of each of the approaches in the list of recommendations
FUTURE WORK
Future Work

• Evaluating the performance of our implementation using a larger set of users in a real environment

• Changing or switch the priority value and the percentage that each of the recommendation approaches and analysing the impact

• Adding a collaborative approach in the recommendation list