UNIVERSITY OF SOUTHAMPTON

A Context-Sensitive Relevance-Based Intelligent Data-Ranking Agent

by

Thomas J. Bell

A project report submitted for the award of
MEng Electronic Engineering

in the
School of Electronics and Computer Science

November 2013

UNIVERSITY OF SOUTHAMPTON

Abstract

School of Electronics and Computer Science

MEng Electronic Engineering

by Thomas J. Bell

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centred vertically so can expand into the blank space above the title too...

Contents

Chapter 1

Project Goals

Recommender systems are forms of information filtering systems which remove unwanted information and predict that which the user is likely to want to read. They can be useful for ranking social media according to a users preferences and context. This report describes an agent for the classifying, scoring and ranking of data according to its user-specific and time-specific relevance. This project description is based upon my initial project brief (Appendix A).

1.1 The Problem

Social media, productivity tools and internet-based information are abundant on mobile devices, leading to users being overwhelmed with information, despite only a small amount of it being of any interest to a particular individual at any given moment. This calls for a means by which such data can be ranked or filtered according to its importance, interest or relevance.

1.2 Project Objective

The objective of this project is to produce a scalable and modular relevance-based ranking agent, to order social media, productivity and other web-based information according to its time- and topic-relevance to a user.

1.3 Goals

The following are core goals which this project sets out to achieve.

1. Develop a scoring algorithm by which to judge to relevance of an item of data based upon a user's
   (a) Personality profile
   (b) Time-related factors
2. To perform automatic remote topic analysis to judge the topic of an item of data
3. Develop a sorting/ranking algorithm to sort or insert scored items of data efficiently, into an ordered list
4. Develop a stable and robust data fusion technique to combine a range of data into a user profile and data profile.
5. Abstract away this agent into an extensible Java API for use in
   (a) Smartphone apps (Android)
   (b) Web-apps (Spring MVC)
   (c) Desktop applications (Java Swing etc.)
6. To develop an application to demonstrate the working API which automatically ranks a user's data according to its relevance

These are the criteria by which the extent of this project's success will be evaluated.

1.4 Unique Features

Recommender systems have been applied to social media before, but never with time-specific relevance factored in. This project is unique in its endeavour to combine recommendation techniques with user- and time-relevant scoring in the development of a commercially viable prototype.

Chapter 2

Background and Report of Literature Review

A progressive project in the sphere of cross-platform relevance-based intelligent ranking agents, using text analysis and a mathematically rigorous scoring algorithms, requires research in a range of areas across the entire spectrum of low- to high-level computational theory and existing product research. The following summarises the research undertaken before and during the research and design phases.

2.1 Existing Data-Ranking Implementations

A good number of content aggregators exist at present in various forms, yet all distinctly lack the complementary relationship between social media (and others) and relevance-based ranking.

Google Now combines Google's search feature with weather and navigation information customised to the user. ViralHeat allows commercial users to filter content from twitter, Facebook and others according to its sentiment (positive/negative), but does no ranking. StreamLife aggregates social media, but performs no ranking or productivity-based data integration.

2.2 Text analytics libraries

There are a good number of existing text analytics services available to the end user and the developer for a range of different types of analytics. These services include sentiment analysis, text categorisation, contextual targeting and a range of others.

Alchemy provides an API which performs sentiment analysis and text categorisation (among others). It provides the core features that this project requires, however the text categorisation often failed to make any categorisation. The best solution in general was Semantria, but is expensive to use.

DatumBox is a free machine learning API which performs sentiment analysis, subjectivity analysis, topic classification, language detection, readability detection, educational detection, document similarity analysis, and gender detection. Many of these features may be useful for ranking text based upon its relevance to a particular individual. It has a simple API using http POST requests and a JSON response.

2.3 Data-mining

I've explored a variety Facebook, Twitter and Android APIs and have written up how to this (see Appendix **??**).

Facebook4J is a Java Facebook wrapper API which simplifies the most common Facebook API features into a more minimal library. Simlar to Facebook, Twitter provides an API which is freely available yet overly complex for this project. Twitter4J is a free API which simplifies the Twitter API.

Calendars from a user's Google account can be retrieved on the Android platform using the Calendar Provider. This is a repository of a user's calendar events which can be queried. Tasks can be retrieved from a Google account using the Google Tasks API.

SMS messages can be received in Android applications using a BroadcastReveiver which collects incoming SMS messages.

2.4 Semantic Analysis

This project requires topic analysis of items of mobile data, in order to compare them to the user's preference and ascribe relevance to them. Other semantic analysis capabilities would prove beneficial for increasing the range of criteria by which relevance may be judged and to eliminate irrelevant data as early on as possible. These include readability, gender, subjectivity and language detection. The DatumBox API is chosen for semantic analysis for its coverage of these requirements; its ease of implementation and the fact that its use is free.

2.4.1 DatumBox API and usage

Each feature provided by DatumBox has a POST Request URL and is retrieved in code by setting up the request headers and URL parameters (including the API key and text), and waiting for the asynchronous response as a JSON object **??**.

2.5 Sorting Algorithms

Sorting is required for ordering scored items of data into a list whereby the topmost items are the most relevant and the bottommost are the least relevant. Since we are only dealing with relatively small sets of data, efficiency is not paramount in terms of maximising accuracy or usability. Despite this, some consideration is made to using the most appropriate sorting algorithms. Merge sort which uses a divide-and-conquer approach is one of the most suitable for initial scoring of larger numbers of items. It is stable and due to its constant performance of $O(nlogn)$, predictable in terms of execution time. For nearly sorted lists such as scenarios where single or small numbers of data items (¡10) may be added to an ordered list, insertion sort is superior in speed and simplicity to others and will have a performance of $O(n)$.

Insertion sort is ideal for the requirements of this project. It is fast for small, largely sorted lists and easy to implement. The following insertion sort algorithm will sort items according the their score (Algorithm 2.5.1).

Algorithm 2.5.1: INSERTIONSORT(A)

comment: Sort the array of items of data D

for $i \leftarrow 1$ **to** $length(D) - 1$

do
- $key \leftarrow D[i]$
- $j \leftarrow i$
- **while** $j > 0$ **and** $score < D[j-1]$
  - **do**
    - $D[j] \leftarrow D[j-1]$
    - $j \leftarrow j - 1$
- $D[j] \leftarrow key$

2.6 Data Sources and Persistence

This project attempts to develop a framework for use in a wide range of software applications and as such will require complete flexibility in terms of its data sources. For the

purposes of research, development and initial testing, JSON (JavaScript Object Notation) objects in text files are considered the best solution for long-term storage of test data. They are simple to create, read, update and delete; flexible in terms of multi-device usability and fully interchangeable with Java objects. Thus they are suitable for storing items of test data to be loaded into the API. UserContext objects may also be stored as such.

Android provides a range of ways to save persistent application data. Data storage options include: shared preferences, internal storage, external storage, SQLite databases or on an external server accessed remotely [3]. Any number of these could be used to store information about the user, such as their preferences and UserContext which is required by the ranking agent. Internal storage will store text files in the file system. They are private the user and other applications and can be used to store JSON objects. A database may be used to store test data and user-related data, but this would require classes for managing a database connection and parsing the data fields into their original objects. Either internal or database storage would be perfectly adequate.

2.7 Recommender Systems

Recommendation systems aid users in finding relevant data and suggesting items (such as Tweets, appointments, news articles) which may be worth reading in more detail. This section explores the two types of filtering for recommendation and principles for exploiting them, with the mobile application particularly in view.

2.7.1 Collaborative Filtering

Collaborative filtering methods [4] use information about the behaviour and activity of various users, and use their similarity to other users to predict which other users will like the same content. It does not rely upon understanding complex characteristics of items and is therefore more accurate for complex items which are hard to machine analyse to extract characteristics. They are, however, dependent upon existing data on a user and difficult to scale. Collaborative filtering systems do not have to understand the item, but only the similarity between users. Pattern recognition algorithms such as 'k-nearest neighbours' is commonly used to measure the similarity between users or items in recommender systems. They are used by applications such as Amazon's recommender, Last.fm and social networks.

2.7.2 Content-based Filtering

Pazzani and Billisus discuss content-based approaches [5] which use characteristics of the items to be recommended to match them to items similar to those the user has liked in the past. Content-based filtering uses an item profile (a tuple of discrete attributes) characterising the item of data. The weight of each attribute denotes the extent to which each feature describes the item and is compared to a profile of the user's interests. The most basic methods use average values of the attributes to denote importance. Most complex systems use Bayesian classifiers, cluster analysis, or decision trees, among others.

2.7.3 Hybrid Recommender Systems

Hybrid recommender systems combine the two approaches by combining the results of each; adding capabilities from one to the other; or by attempting to unify the ideas behind both into a single model.

Due to the complexity of reasons why users may find an item relevance, many modern systems [6] use collaborative filtering to incorporate social relationships. Sen et al. [7] among others use tags within content-based approaches to generate better rankings.

2.7.4 The Utility Matrix

In typical recommendation systems a *utility matrix* is a matrix which gives the user-item pair for each user's preference of each item. This is more common in collaborative filtering approaches where a user's recommendation of an item is based upon other users who viewed it.

In content-based approaches, it specifies the preference of the user, of each characteristic or attribute about a particular type of item, such that the item is then classified and scored by comparing the item profile with the user profile/utility matrix.

2.7.4.1 Populating the Utility Matrix

The Utility Matrix is required to be accurate for accurate recommendations and there are two primary approaches in populating them with their relevance to the user. These may be

1. asking the users to rate items/attributes explicitly, or

2. inferring the user's preferences implicitly from their behaviour

Chapter 3

Report on Technical Progress

This chapter highlights my progress so far, by explaining my theoretical findings and highlighting my implementation progress.

3.1 Research Progress

I've chosen to model a user's preferences as a tuple U of attributes u_k which given the users preference for each topic. This is the user profile.

In order to rank data according to its relevance, it must first be scored according to its relevance. Scoring uses a recommendation system in which an item of data is allocated a tuple of attribute scores. These are compared to the users profile to given the item a score.

Using the unweighted scoring rule

$$f(D) = \frac{\sum_{i=1}^{i} d_i}{n} \tag{3.1}$$

where D is our item of data, and the elements d_i are its topic-attributes, I used Fagin and Wimmers' [9] conversion to a weighted rule to give

$$f_U(D) = \left(\sum_{n=1}^{M-1} n \centerdot (u_{\sigma(n)} - u_{\sigma(n+1)}) \centerdot \frac{\sum_{i=1}^{n} d_i}{n} \right) + M \centerdot u_\sigma(M) \centerdot \frac{\sum_{i=1}^{n} d_i}{n} \tag{3.2}$$

$$= (u_1 - u_2)d_1 + 2(u_2 - u_3)\frac{d_1 + d_2}{2} + 3u_3\frac{d_1 + d_2 + d_3}{3} \tag{3.3}$$

$$= u_1 d_1 + u_2 d_2 + \cdots + u_M d_M \tag{3.4}$$

Here $X \upharpoonright \{\sigma(1), \ldots, \sigma(i)\}$ is a restriction of X to the domain of a bijection σ which orders the weightings to match the order of entries in the tuple. In our case this bijection would be a mapping from each attribute of an item of data to the weighting associated with that attribute. The weighting Θ would be dependent upon a complementary tuple describing the user, i.e. their attribute preferences. The term $(\theta_{\sigma(i)} - \theta_{\sigma(i+1)})$ is the difference between the weightings of two consecutive entries.

This lead me to the proof of an intuitive linear weighted scoring rule (Eqn. 3.5).

$$f_U(D) = \sum_{k=1}^{m-1} u_k d_k \tag{3.5}$$

As items are received, they are scored and ordered, but an item may be relevant in terms of topic, but not in terms of age since its creation. It may have been created a long time ago and thus needs removing to make way for new items. Here, I add an item-removal term $\mathrm{e}^{-\alpha t(d)}$ where $t(d)$ is the minutes lapsed since the receiving of item d and α is a delay coefficient. As time increases, the item-removal term tends to 0. This give us

$$f_U(D) = \mathrm{e}^{-\alpha t(d)} \cdot \sum_{k=1}^{m-1} u_k d_k \tag{3.6}$$

Time-relevant data is scored differently to topic-relevant data. Time scoring may be done using a continuous increasing term who's value is related to the time before a deadline approaches. $t_{threshold}$ is the time at which time-related scoring begins, $t_{tillDue}(D)$ is the time until item D is due and β is a normalisation parameter. It must be set such that at the point at which D becomes relevant in time, $f_{time}(D)$ is greater than the maximum non-time-related score. This gives us

$$f_{time}(D) = \mathrm{e}^{\beta(t_{threshold} - t_{tillDue}(D))} \tag{3.7}$$

This gives us a weighted scoring function incorporating both topic-relevance and time-relevance into a single equation (Eqn. 3.8).

$$f_U(D) = \left[\mathrm{e}^{-\alpha t(d)} \cdot (1-\gamma) \sum_{k=1}^{m-1} u_k d_k\right] + \gamma \mathrm{e}^{\,(t_{threshold} - t_{tillDue}(D))} \tag{3.8}$$

3.1.1 Lexical Analysis and Preprocessing

The user profile may also include requirements for the filter to remove items which fall under certain criteria. These include readability, sentiment, spam removal, adult content removal, language detection and gender detection. The preprocessor will eliminate items which are excluded by the user profile, before being scored.

3.2 Implementation Progress

In terms of implementing the algorithm as an API, I've created a test environment for entering test data from the command line and storing them in files as JSON objects. It loads them from files into DataItem objects to be used by the ranking agent. It allows me to chose which types of items are loaded at any point.

I have implemented classes which sort items of scored data; get the topic (among other features) of items; store the users profile; store the item's attribute characteristics and score items according to how well they match the users profile.

Chapter 4

Plan of Remaining Work

4.1 Remaining work

In terms of work remaining, the vast majority of my research and algorithm development has been completed. Only as testing progresses and inevitable changes are be required, will there need to be more research undertaken.

4.1.1 Algorithm and Java API Implementation

The algorithm and mathematical foundation for this project has been established and explained. I've completed a significant amount of the Java API which implements the algorithm and also classes which deal with data persistence, sorting and topic classification. I have yet to test the API rigorously and make alterations to it which may be required for its integration with the demonstration app.

4.1.2 Demonstration Application

The demonstration application has not yet been implemented. This is the main task of the coming term. I need to finish testing and refining the Java API until it works as expected. A skeleton user interface will be made separately from the API and simultaneously with the API testing. The application will then integrate with the API and I'll first test the algorithm using fake test data. Once this is working I'll use real data sources such as Twitter and Facebook accounts, and tasks and appointments from the device.

4.2 Planning and Contingency

After returning from Christmas there will be three weeks of revision and exams, after which 15 weeks remain. I've allocated generous slots for the completion of each of the remaining tasks, and two weeks contingency time before the deadline.

4.2.1 Algorithm and Java API Implementation

The algorithm and Java API has been largely implemented, but significant amounts still remain. I've allocated a remaining 11 weeks in total for the completion of the API and the demonstration app, to allow for testing and the write up of the final report before the deadline. Over the Christmas break and shortly afterwards, I will be completing the API and validating its results. In terms of contingency planning, should this process overrun significantly, it may be required that a console or desktop demonstration application should be used as opposed to a prototype mobile application.

4.2.2 Demonstration Application

From early February, the development of the mobile application will begin. This will run alongside the write up of the final report, and continued testing will inform the validation process throughout, such that before the Easter before the deadline, an entire system will have been researched, designed, implemented and tested. This will leave time for a thorough analysis and evaluation of my findings, and for though concerning future work.

Tasks	Start	Duration (Days)	Complete (%)
Literature Review	**30/9/13**	**84**	**70%**
Existing Implementations	30/9/13	7	100%
Data Acquisition	30/9/13	14	100%
Semantic Analysis API's	30/9/13	14	100%
Research scoring	30/9/13	70	75%
Research sorting	18/11/13	28	80%
Algorithm Design	14/10/13	63	60%
Weighted Scoring rule	14/10/13	42	85%
Time-relevance term	19/11/13	28	50%
Implementation	**1/11/13**	**126**	**40%**
Ranking Agent API dev.	1/11/13	70	60%
Demonstration App dev.	N/A	63	0%
Validating Results	**N/A**	**84**	**0%**
Compare with manual	N/A	42	0%
Find match statistics	N/A	70	0%
Interim Report Writing	**N/A**	**77**	**100%**
Write up project brief	3/10/13	14	100%
Write up literature review	3/10/13	63	100%
Write up progress so far	25/11/13	21	100%
Write up work still to do	26/11/13	7	100%
Final Report Writing	**N/A**	**112**	**0%**
Update literature review	N/A	14	0%
Write up algorithm design	N/A	14	0%
Write up design and impl.	N/A	49	0%
Write up analysis and eval.	N/A	42	0%
Exams/Holiday	**N/A**	**21**	**0%**

Appendix A

Project Brief

Problem Social media, productivity tools and internet-based information are abundant on mobile devices leading to users being overwhelmed with information, despite only a small amount of it having any interest to a particular individual at any given moment.

Goals The goal of this project is to provide a generalised ranking agent to order this vast range of information according to its context-specific relevance, given the user's personality, click-history and environment at any instance. This project will endeavour to abstract away a new extensible ranking agent into a Java API for use in a variety of applications on a range of devices.

Existing ranking, sentiment analysis and data fusion algorithms will be investigated, employed or adapted in order to produce a scalable and highly modular context-sensitive mobile-content relevance-based ranking agent.

A personality profile and historical data will be used to maintain a user-context, which will behave like a search query. Topic analysis will be used to determine the data-context of a variety of available items of data, to be matched for relevance against the user-context. A stable context-sensitive ranking algorithm will be proposed and implemented to order data according to its context-specific relevance.

The Facebook and Twitter API's will be investigated to ensure data-object compatability and the Android, Spring MVC and Java Swing frameworks will be explored, to ensure compatability with Java mobile, web-based and desktop applications respectively.

Realistic test data will be used throughout the development stages in a comprehensive variety of configurations. The primary deliverable will be a modular Java API. For the

purpose of demonstration a mobile, web or desktop application will be designed and implemented, to showcase the ranking agent's capabilities.

Scope For the purpose of allowing this project to focus on its main goals, it will not intend to improve upon existing sentiment/topic analysis algorithms at the outset, but rather use existing tools.

Compatible items of data within the initial scope of this project include Facebook statuses/notifications, tweets, calendar appointments, tasks, emails and SMS messages and will allow for additional types of data to be added later.

The project will focus primarily on the ranking agent, designed to rank the items of data using modified algorithms on a specific set of feature vectors.

Bibliography

[1] Facebook4j website. URL `http://facebook4j.org/en/code-examples.html`.

[2] Twitter4j website. URL `http://twitter4j.org/en/code-examples.html`.

[3] Wei-Meng Lee. Beginning android 4 application development. pages 263–291, 2012.

[4] B. Oki D. Goldberg, D. Nichols and D. Terry. Using collaborative filtering to weave an information tapestry. *ACM35*, pages 61–70, 1992.

[5] Michael J. Pazzani and Daniel Billsus. Content-based recommendation systems. *The Adaptive Web*, pages 325–341, 2007.

[6] Henry Kauts, Bart Selman, and Mehul Shah. Referralweb: Combining social networks and collaborative filtering. *ACM40*, pages 61–70, 1997.

[7] J. Vig S. Sen and J. Riedl. Tagommenders: Connecting ysers to items through tags. *WWW 2009*, pages 61–70, 2009.

[8] S E Robertson, S Walker, S Jones, M M Hancock-Beaulieu, and M Gatford. Okapi at trec-3. *NIST*, 1995.

[9] Ronald Fain and Edward L. Wimmers. A formula for incorporating weights into scoring rules. *Elsevier*, 2000.