University of Bergen

Master Thesis

Personalization of tourist application using semantic technologies

Author:
Lisa Halvorsen

Supervisor:
Csaba Veres

in the

Departent of Information Science and Media Studies

June 17, 2013

UNIVERSITY OF BERGEN

Abstract

Faculty Name

Departent of Information Science and Media Studies

Master

Personalization of tourist application using semantic technologies

by Lisa Halvorsen

The main research question this thesis tried to answer was: "Using semantic technologies and information collected from a user's social network profile, is it possible to generate a reliable model of that user's interests?" Some research has been done using semantic technologies to create user models, and social networks have been used to collect information about the user's interests in order to apply that information to recommender systems. This project however contributed to the field by investigating the combination of using Facebook as a source for the user's interests, and using semantic technologies (topic modelling and RDF modelling) of that information to create a user model which will be applied to a different domain. Tourist recommendations were chosen as the other domain because of personal love of travelling and problems with finding the right kind of information about new destinations. A prototype Android tourist application was developed to demonstrate the concept. The conclusion of the project was that it is possible to create a reliable model of the user's interest using topic modelling and RDF-modelling of the user's Facebook information. There was however potential for improvement in applying this user model to the tourist domain.

Contents

List of Figures

List of Tables

I would like to thank all who have helped me during the project. Especially I would like to thank my supervisor Csaba Veres who has been a great help and motivator. I would also like to thank all the persons who helped me test the Semantic Tourist.

Another big thanks to fellow students, especially my guys in room 635 and my sister Mari. You have motivated, entertained and made the last year memorable. I also thank my family and friends for the support.

Most of all I would like to thank my great boyfriend Max for all help. I hope you know how much I appreciate your encouragements and a patient mind.

Chapter 1

Introduction

"*I planned on going to Thailand in December, however San Francisco seemed so exciting that I am considering going there instead*". - Comment from one of the test participants comments during the test session of the Semantic Tourist.

A problem with traditional tourist guides is that they are too general. They have to cover "everything", and focus on what they presume the "average" tourist will find interesting. I wanted to develop a tourist application that will recommend the things *I* personally want to see and do. There are several ways in which this can be achieved. User modelling of a each single user is one of them. The user model should contain information about the user's interests and personal preferences (Baldoni et al., 2005). One of the problems with applying user models in recommender systems is in how the system gathers enough information about the user. Systems, such as the one presented by Aek (2005), creates a user model by collecting information from the user's past behaviour, while other approaches, such as Burke (2000) use domain ontologies to help the user enter information about what they are interested in. One problem with the first approach is that the system has to be used for a minimum amount of time before it has gathered enough information to come up with recommendations. This is called the cold-start problem (Maltz and Ehrlich, 1995). A problem with the latter approach, is that it is restricted by the guidelines the system uses in aiding the user-input with, and the information the user enters into the system. It is not certain that the system will be able to find all the things that the user might want to see and do. Middleton et al. (2004) claims it is easier for people to recognize what they want once you present it to them, rather than asking them to articulate what they want. In order to present the user with all the things that he or she would potentially want to see or do, the system has to figure our what the user likes. If the system knows the user's interests and preferences, not just in the tourist domain, it could use the information to infer things that the user potentially would like to do while on vacation.

1.1 Research project

The idea behind this project was to use the user's social network presence, to learn what he or she might like, because people share and express their interests and opinions here. There have been some studies on using information from the users social network to leverage the cold-start problem. However they have mainly focused on the users' connections to compute similarities between users. One study has attempted to extract the user's published content on Facebook. Facebook was chosen as primary source for this project because it has a rich user profile and a large user group. However the system should be designed in such a way that other data sources will be easy to add and integrate in the future. The system uses semantic technologies to create an individual user model of the user's preferences. The combination of source for the user models and the technologies used to created the models is the innovative part of the project. The created artefact will try to demonstrate new take on solving the "old" problem of giving personalized content-based recommendations.

1.2 Goals

The broader goal of this project is to investigate if it is possible to generate a model of the user's interests using his or hers social media presence and semantic technologies. More specifically, to investigate if, it's possible to extract enough semantic information from the user's Facebook profile to generate a reliable user model of the user's interest that then can be used to generate recommendations in the tourist domain.

The project will create a system prototype that uses semantic technologies to collect information from the user's Facebook profile. This information will be used to generate a model of the user's interests. There will be an investigation of the information to see if parts of the profile information are more useful than others. The tools utilized will also be investigated to see if it is possible to fine tune them in order to achieve better results. To test the reliability of the user model it will be investigated whether the users recognize it as a representation of their interests, and of their Facebook presence.

The created user model will be used as a base to generate personalized recommendation in the tourist domain. The recommendations are given in points of interest (POI) where a point of interest is a location, attraction, event or similar that is of interest to the user. Interest in this context is that the user would consider visiting the point of interest if he or she were to visit the city.

The evaluation of the system will be a quantitative measure of the quality of the user model applied to the tourist domain. It will be followed by a qualitative questionnaire to further understand and explore the quantitative results.

The scope of the prototype developed for the project is to combine exciting libraries in a prototype application that can be used to test the concept. It includes a simple client application demonstrating the features. However good usability, user session handling, security and other features related to publication of the application is outside the scope. The thesis will explain about the methods used in this work and field specific terms and methods. However basic software engineering terms and methods will not be explained in detail.

1.2.1 Research questions

Based on the goals described above the following research questions are proposed:

- Using semantic technologies and information collected from the users social network profile, is it possible to generate a reliable model of a user's interests?
- Does Facebook provide sufficient information to do this?
- To what degree can the resulting model be used to generate recommendations in the tourist domain?

1.3 Personal motivation

This project is motivated by personal experiences when looking for things to do while on vacation. I really enjoy travelling, but I'm not too fond of all the museums, art galleries and other typical tourist attractions that are often recommended when browsing guidebooks or the Web. When I travel I want to experience the place I'm visiting like a local person. What I want from the optimal tourist application is for it to show me hidden restaurants where only locals go, activities that I can take part in like diving, skiing, rafting, skydiving etc. and events related to my interests in sports, food and programming, among others.

1.4 Organization of the thesis

The thesis is organized as follows; Chapter 1 provides an introduction to the thesis and a presentation of the research questions the thesis will try to answer. Chapter 2 presents the theoretical framework for the thesis. Chapter 3 presents the research framework and the software development framework this project has followed. The fourth chapter describes and explains about the components of the Semantic Tourist application. Chapter 5 presents the evaluation of the application conducted during its development. It also contains details on the final evaluation. The results of the final evaluation are presented in the first section of chapter 6, while a discussion of the results with an eye to answering the research questions is presented in the second section. The final chapter, Chapter 7, contains the conclusion of the thesis and suggestions for further research on the subject.

Chapter 2

Theoretical framework and literature review

This chapter will present the theoretical framework that forms the foundation of the research project. The first part will be a general introduction to semantic Web, the main research field of the thesis. It will explain concepts like Linked Data, RDF-model, ontology and SPARQL. Then a presentation of what personalization is and how it can be accomplished follows. After this a more detailed explanation about user modelling and recommender systems is presented. The chapter finishes with a presentation of what topic modelling.

2.1 Semantic Web

The World Wide Web is a system of documents linked together through hyperlinks accessible through the Internet (Hebeler et al., 2009). Text, images and videos etc. are represented in Web pages which are accessible to humans through Web browsers. The content of the Web pages is normally not understood by computers. The purpose of the world wide Web was to make documents available in a format that can be read by humans. But there is a downside to this. Computers do not understand the documents. They can follow hyperlinks and display the content of the page in a way that humans can read and interpret the content. Computers facilitate navigation between documents, but a human must check that the content satisfies the information need.

The semantic Web is an additional layer to the World Wide Web which is sometimes refereed to as the Web of Data (Bizer et al., 2009). The Web of Data is data on the Internet that is published in such a way that it “is machine-readable, meaning it is explicitly defined, it is linked to other external data sets and can in turn be linked to from external data sets” (Bizer et al., 2009). This allows computers to navigate between the data on the Web, using links between data, in a way similar to how humans brows the Web following hyperlinks.

Linked data is a set of rules of publishing data on the Web that provides the means to contribute to the creation of the semantic Web. These rules are known as the “Linked Data principles” and are as follows (Berners-Lee, 2006):

- Use URI’s as names for things
- Use HTTP URIs so that people can look up those names
- When someone looks up the URI, provide useful information.
- Include links to other URI’s, so that they can discover more things.

Uniform Resource Identifiers (URI) and HyperText Transfer Protocol (HTTP) are two fundamental technologies in the Web that Linked Data relies on (Bizer et al., 2009). URIs is a super set of the more familiar Uniform Resource Locators (URL). While URLs address documents and other entities on the Web, URIs are more generic and can identify any entity that exist in the world. URIs that use the `http://` scheme can be looked up over the HTTP protocol. HTTP provides a mechanism for retrieving resources that can be serializes as stream of bytes (e.g. a picture of the mountain ”Fløyen”), or descriptions of things that can not be sent over the network (e.g ”Fløyen” itself).

Returning to the earlier discussion of the Web of Data; for the data to accessible by computers, it needs a common structure across different data sets and a way to link between them. The Resource Description Framework (RDF) is one of the technologies that can be used to achieve this. It is a framework for modelling the data. In RDF everything is expressed as triples. A triple consist of a subject, a predicate and an object. This is modelled as a graph where the subject and object are nodes, and the predicate is the edge between the nodes. There are two kinds of nodes: resources and literals. Literals represent all concrete data values like strings and numbers. Literals can only be the object, never the subjects in a triple. In contrast to the literals, the resources can represent anything else, and they can be both subject and object. A resource can represent anything that can be named - an object, act or a concept. The resource takes the form of a URI. The predicate, also called property, is also a resource which connects the subject and the object (Hebeler et al., 2009).

Listing 2.1 shows an example RDF statements about Bergen. The first states that the resource identified by `http://example.com/#Bergen` is a type of the other resource `http://example.com/#City`. If the example.com resources were real you could look them up using the HTTP protocol. You cold see that `http://example.com/#Bergen` was a page about Bergen, `http://example.com/#City` was a description of what a city is, and `http://www.w3.org/1999/02/22-rdf-syntax-ns#type` tells you that "a resource is an instance of a class" meaning Bergen is an instance of the class City (W3C, 2004). The second triple statement connects the example description of Bergen to the DBpedia description of Bergen. It uses the Web Ontology Language (OWL) which is a semantic markup language for publishing and sharing ontologies on the Web. The statement uses "owl:sameAs" to link the two individual instances of Bergen. The "owl:sameAs" is used to state that two URIs reference the same thing. This statement serves as a link between the example data source and the DBPedia data source (Bechhofer et al., 2004).

```
Subject: http://example.com/#Bergen
Predicate: http://www.w3.org/1999/02/22-rdf-syntax-ns#type
Object: http://example.com/#City

Subject: http://example.com/#Bergen
Predicate: http://www.w3.org/2002/07/owl#sameAs
Object: http://dbpedia.org/page/Bergen
```

Listing 2.1: Example RDF statements.

RDF graphs are well suited for making humans understand the representation of the information. It is however not practical for information exchange between computers. To solve this problem there are different forms of RDF serializations. These are ways to convert between the model and a concrete data format. The three most popular are RDF/XML, Terse RDF Triple Language (turtle) and N-Triples (Hebeler et al., 2009).

To create more meaning, or rather add more semantics to the data, an ontology can be created on top of the data sets. An ontology is a "formal and explicit specification of a shared conceptualisation" (Gruber, 1993). W3C (2013) states that "there is no clear division between what is referred to as "vocabularies" and "ontologies"", and states that "vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms."

An ontology can also link data sets. It can be created with ontology languages like OWL and RDF Schema (RDFS) (Bechhofer et al., 2004) (W3C, 2004). The Semantic Tourist make use of the owl:sameAs property in the Linked Geo Data. With OWL you can for instance say that "Bergen" from your model is the same as the location "Bergen" from LinkedGeoData and the same as the resource "Bergen" in DBPpedia. The ontology acts as a connector between the data sources, and enables the possibility of combining data. An example demonstrates this better. You are looking for all the tourist attractions in Bergen. Through LinkedGeoData (LGD) you can find all locations in Bergen. One of the retrieved results is Fløyen, but LinkedGeoData only provides geographical data and states that it is a tourist attraction (LinkedGeoData.org, 2013).DbPedia provides more information about Fløyen. It has descriptions, information about how high the mountain is, pictures of it, the name of the area it is located in etc. (Dbpedia.org, 2008). By linking LDG resource with the DBPedia resource through e.g a owl:sameAs property it is possible to follow the link between the two dataset.

The SPARQL query language is used to query the Linked Data. Unlike traditional SQL, which is used to retrieve data from relational databases, SPARQL queries in terms of graphs and triples. Listing 2.2 show a example query retrieving the name and the population of the city which is the capital of Norway. The PREFIX ex:<http://example.com/> states that all instances of "ex:" is a short version of `http://example.com`. Variables are notated with a "?". So select name and population from the model, where "?x" is the subject, and "?x" has the properties cityname, population and capital and the object in the triple "?x ex:capital ex:Norway" is "ex:Norway"(Prud'hommeaux and Seaborne, 2008).

```
PREFIX ex:<http://example.com/>
SELECT ?name ?population
WHERE {
        ?x ex:cityname ?name .
        ?x ex:population ?population .
        ?x ex:capital ex:Norway.
}
```

Listing 2.2: SPARQL query example. It selects the name and the population of the city which is the capital of Norway.

?name	?population
Oslo	613,000

Table 2.1: Example result from the SELECT query in Listing 2.2

There are four ways of querying with SPARQL: SELECT, CONSTRUCT, ASK and DESCRIBE. The SELECT query returns the result as variables and bindings. The response of the SELECT query in Figure 2.2 is listed in Table 2.1. The CONSTRUCT query returns an RDF graph specified by the template in the query. It takes the response from each query solution, substitutes the response with the variables in the template and combines the triples to a single RDF graph. The ASK query simply returns true or false depending on wether the query pattern exists or not. The DESCRIBE query also returns an RDF graph, but unlike the CONSTRUCT query, the response graph is not prescribed by the query. The query is used to create the result, but instead of returning only variables like in the select query, the SPARQL query processor creates a graph describing the result resource. Typically a graph containing the requested resource with all its attributes is returned (Prud'hommeaux and Seaborne, 2008).

2.2 Personalization

The concept personalization means different things to different people in different fields (Fan and Poole, 2006). It can be the recommendation of a hotel after the user's preferences, a gift created for one specific individual or individual mapping of short cuts in a program. The areas of marketing/e-commerce, computer science/cognitive science, architecture/environmental psychology, information science, and social sciences including sociology, anthropology, and communication all have different definitions of the concept(Fan and Poole, 2006). "Personalization is a toolbox of technologies and application features used in the design of an end-user experience" (Of and Acm, 2000), is one example of the definition from computer science. Information Science define it as "Fine-tuning and prioritizing information based on criteria that include timeliness, importance, and relevance to the audience" (Bender, 2002). A more general definition of the concept is that personalization is "a process that changes the functionality, interface, information access and content, or distinctiveness of a system to increase its personal relevance to an individual or a category of individuals"(Blom, 2000).

There are three dimensions to personalization: *what* to personalize, *whom* to personalize, and *who* does the personalization (Fan and Poole, 2006). What to personalize represents what part of an information system that can be changed or manipulated to make it more adapted or relevant to the user. It is distinguished by four aspects that can be personalized. One aspect is the information itself (*content*), the second is what the user can do with the system (*functionality*), the third is how the information is presented (*user interface*), and the forth is through what media the information is presented to the user (*channel*). Only content personalizations will be presented in more detail, as this is the type of personalization utilized in this project. See Section 2.4

Fan and Poole (2006) second dimension is whom is the target of the personalization. They distinguish between a *category* of individuals and one *specific* individual. A category of individuals can be a family, women, single people etc. This type of personalization is also sometimes called stereotype personalization (Rich and Sciences, 1983). If the system is targeted to these categories of individuals, and one individual identifies with this group, the individual will feel that the content is personalized to him or her. Individual personalization is the other option for whom to target the personalization. In these systems the content, functionality, interface or channel is adapted and unique to each single individual user.

The third dimension is about how the information base in the personalized system is created (Fan and Poole, 2006). One option is to make the user enter the information *explicitly* through guidance by the system. The other option is that the system automatically collects the information. This is called *implicit* personalization. The second and third dimensions will be be presented in more detail in Section 2.3

2.3 User modelling

User models should contain information about the user. It can contain general information like name, age, gender, location etc. or more specific characteristics about the person, for example their interests and personal preferences (Baldoni et al., 2005). Another type of user model contains information more related to the activity/thing which the user model is used for. This can be about the user's goals, in the context of the domain or application.

According to Rich and Sciences (1983) there are three important characteristics of user models:

- One model of a *single*, canonical user vs. a *collection* of models of individual users.

- Models *explicitly* defined either by system designer or by the users themselves vs. model *inferred* by the system on the basis of the user's behaviour.
- Models of fairly long-term characteristics such as areas of interest or expertise vs. models of relatively short-term user characteristics such as the problem that the user is currently trying to solve.

The first characteristic is whether every user is modelled individually or if the model represents a stereotype or a class of people (Kabassi, 2010). The downside with the modelling of an *individual* user, is that it requires a certain amount of information before the user model can be used. Hence this method suffers from what is called the cold start problem (Maltz and Ehrlich, 1995). This is when the system does not contain enough information about the user to be able to give recommendations. It is also called new-system-cold-start problem. There is also another type of cold-start problem (new-user) related to collaborative recommender system approaches, which occurs when a new item is added to the system and there are no ratings of the item to base the recommendation on (Middleton et al., 2004). The *stereotype* modelling on the other hand allows for creating recommendations with the first interaction. The downside is that the stereotype can fit the user in some characteristics, but be totally wrong in other. Another disadvantage is that users characteristics may change over time.

The second characteristic is based on the way the data is acquired (Rich and Sciences, 1983). There are two ways of doing this, either ask the user to explicitly provide the data, or attempt to infer it implicitly from other data. In the first case the user has *explicit* control over the information he or she provides to the system. This can be negative because the user may not be able to describe their preferences accurately, which will give an inaccurate model which in turn will result in less accurate recommendations (Rich and Sciences, 1983). Another downside to this approach is that it might be a tedious process for the user to enter the information the system needs to be able to give personalized predictions (Rich and Sciences, 1983). The other approach is one where the system collects the user's information from the user's previous activities, which requires little effort by the user. In this *implicit* approach the user can not give wrong or leave out information, but depending on the information collected the user might have issues about sharing it. They might feel a violation of their privacy (García-Crespo et al., 2009).

The third characteristic involves short-term vs. long-term user modelling (Rich and Sciences, 1983). With this characteristic, in contrast to the other, there is not a more preferable way

way of modelling the user. For a recommender system to work optimally it is important for it to have a wide variety of information about the user. This can range from the long-term facts about the user's general interests to the short-term facts about the goal of this particular vacation. The more responsive the system is meant to be, the more short-term information is needed.

There has also been some research in user modelling using different semantic technologies. User profiles are typically built based on knowledge about the user or the behaviour of the user (Middleton et al., 2004). The former approach uses a static model of the user which dynamically matches the user profile to the closest model. To obtain information for these user models questionnaires and interviews are often utilizes. The latter way to create a model of the user is to collect information from the user's past behaviour (Middleton et al., 2004). The behaviour is logged and machine-learning techniques are commonly used to discover patterns in the user's behaviour. This is the most common to use in recommender systems.

In his study "Ontological User Profiling in Recommender systems" Middleton et al. (2004) presents the two systems Quickstep and Foxtrot. They are on-line recommender systems for academic research papers which uses implicit monitoring of the users behaviour and relevance feedback to create a user profile ontology. These ontologies are used in a collaborative recommender system to avoid the cold-start problem. The study shows that pairing ontologies that share similar concepts had significant success. Garlatti and Iksal (2003) is another study that uses ontologies. In this research four ontologies are used to personalize an adaptive hypermedia environment which manages selection, organization and adoption at knowledge level.

Aek (2005) uses implicit user modelling through an information scent algorithm presented by (Pirolli and Card, 1995) The algorithm sniffs around looking at the how the user navigates through Web pages. It uses "information scents" to create an ontology of the visited pages and derive a user model containing words like ski, snowboard, mountain and winter to generalize that the user likes winter sports. It improves the system by letting the system understand the visited content at a semantic level.

2.4 Recommender systems

"People find articulating what they want hard, but they are very good at recognizing it when they see it" (Middleton et al., 2004).

Personalized tourism services aim at helping the user find things they might be interested in, easily, without spending much time and effort. A variety of approaches have been used in this domain: content-based, collaborative, demographic, knowledge-based or hybrid (Montaner, 2003). Content-based recommendation and collaborative filtering are the two most popular methods. In content-based recommender systems items are chosen based on analysis of what actions the user has taken in the past (Middleton et al., 2004). The system matches the characteristics of items or services against the characteristics of what the user preferred in past. The advantage of this is that the recommendations are based on facts. But this can also be a disadvantage when the recommended item is too similar to what the user already has done. Nothing completely new will be recommended by strict content based recommender systems. Another problem with this approach is first time users. The system has no information to build the recommendations on. This problem though, can be overridden with the addition of other techniques, like stereotyping.

Collaborative filtering on the other hand, is when the recommendation is based on ratings from the users. Collaborative approach has the advantage that people give ratings and with more ratings it becomes more likely that the system will find a good match for the new user. Another advantage is that the system can identifying cross genre niches and can entice the user to jump outside the familiar (Burke, 2002). However, these systems must be initialized with large amounts of data before it can give good recommendations. Collaborative recommender systems often suffer from what is called the "ramp-up" problem: "until there is a large number of users whose habits are known, the system cannot be useful for most users, and until a sufficient number of rated items has been collected, the system cannot be useful for a particular user"(Burke, 2000).

Shapira et al. (2012) uses information collected from the user's social network to supplement and/or replace user ratings in a collaborative recommender system. They extracted the information the user had published on their personal pages about their favourite items and preferences. This information was the used in several experiments; to give recommendations in the same domain and cross-domain, replacing recommendation by rating and supplementing it. They used a similarity measure of the user's published e.g movie items and her friends moive items, and returned the items published by friends with a Jaccard similarity of above 0.5. In the cross-domain experiment they used the user's information in various domains (movies, music, books etc.) and recommends items in one domain based on the similarity with the friends' preferences in all domains. They compared the performance with traditional collaborative filtering methods. They tested 95 subjects and the results show that when data is sparse or not available for a new user, recommendation results relying solely on Facebook data are at least equally as accurate as results obtained from user ratings. The experimental study also indicates that enriching sparse rating data by adding Facebook data can significantly improve the results.

The knowledge based recommendation system uses knowledge about the user and the item to generate recommendations based on reasoning about what items meets the users requirements. One advantage over the two others is that these systems do not suffer from the "ramp-up" problem because it does not depend on user ratings. Additionally they do not have to gather information about particular users because they give recommendations independent of the user's individual taste. The drawback of these systems is the comprehensive knowledge engineering (Burke, 2000). Burke (2000) uses a knowledge-base for recommending restaurants where the recommendations is given based on one known item. The system recommends items that are similar to an "item x". Towle and Quinn (2000) propose a knowledge-based system that allows the user to critique the recommendations made by the system. This approach is more focused on what the user likes/needs. Michelson and Macskassy (2010) use Wikipedia as a knowledge base to leverage the disambiguation in the content of the Tweets. A similarity measure of the context, the other words in the tweet, is compared to the category system in Wikipedia to disambiguate each term. Their experimental study show promising results.

Kabassi (2010) gives an overview of the research that has been done on personalizing recommendations for tourists. The article presents the different approaches to user modelling and recommender systems and research related to various concepts. She lists three studies about systems recommending points of interests based on interest/goals, and five which combine individual user modelling and implicit information acquisition.

In "The mobile Wine Agent: Pairing Wine with the Social Semantic Web" Patton and Mcguinness (2009) developed a semantic Web application for making wine and food recommendations to users. The application uses data from an underlying ontology to drive the user interaction and give recommendations. Social applications like Facebook and Twitter are utilized to share content with other users of the world wide Web and to make the recommendations personalized through the user's shared content in the network.

In "Personalized social search based on the user's social network" Carmel et al. (2009) investigates personalized social search based on the user's social relations in social networks. The social network is used as a basis for familiarity-based recommendations (friends), similarity-based recommendations (social activities) and an overall approach which includes both. These approaches are versions of collaborative recommendations. Their results outperform the non-personalized and the topic-based approaches that they compare with.

2.5 Topic modelling

Earlier we talked about the semantic Web and how this is data integrated in the Web so that the computer is able to comprehend the data. Semantic technologies are what makes this possible. However, if there is no or little structure to the information source you need a different approach to make the computer understand the information. There are various approaches; Elberrichi et al. (2008) used wordNet to categorize text, Hotho et al. (1998) used ontology aided clustering techniques, and there are other machine learning approaches. Only topic modelling will be presented in more detail, as it is the technology used in this project.

Topic modelling can be used to figure out what a document or a collection of documents is about (Blei et al., 2003, Oh, 2010). The purpose is to uncover and understand the underlying semantic structure of the document(s). The Latent Dirichlet Allocation (LDA) is the simplest, and most common topic model (Blei et al., 2003). It is a generative model. This means that it tries to mimic what the process of writing a document is. It tries to generate the documents given the topics. "A topic is a distribution over terms in a vocabulary"(Blei et al., 2003). The vocabulary is all the words in the collection of documents. Each document in the collection belongs to a topic with some probability. The topic includes all the words in the vocabulary, but the words with highest probability of belonging to the topic will describe the topic best. The distinguishing characteristic of the Latent Dirichlet Allocation compared with other topic models is that "all documents in the collection share the same set of topics, but each document exhibits those topics with different proportion"(Blei, 2011). The following example will explain this in more detail (Oh, 2010).

An article is about three topics: NASCAR races, economic recession and general sports topic (see table 2.2) (Oh, 2010). The words in the right column are words that have a high probability of being in the corresponding topic. Actually all words in the article are a part of the topic, just with a higher or lower probability of belonging to the topic. If you want to generate an article about these topics you would have to figure out the distribution of the topics. So say the article

NASCAR races	track, raceway, cars,
Economic recession	sales, cost, business
General sports	athlete, competition, physical, activity

Table 2.2: Example topics in an article and words form the article which are associated with the topic.

is mainly about NASCAR. When writing about it you would use the words corresponding to that topic more than the other words. This is how the document is generated according to the assumption of the LDA. The documents in the collection are observed, while "the topic structure - the topics, per-document topic distributions, and the per-document per-word topic assignments - are hidden structures" (Blei, 2011). Hence when you have an actual document or document collection, the work of LDA is to use the provided document collection to infer the hidden topic distribution. See the original article "Latent Dirichlet Allocation" (Blei et al., 2003) for further details.

Some research has been done applying topic modelling to social networks. In "Investigating Topic Models for Social Media User Recommendation" Pennacchiotti and Gurumurthy (2011) use topic modelling to recommend to a user new friends who have similar interests to their own. They use the users' social media streams from Twitter to represent their documents in a adaptation of the LDA algorithm. Furthermore they represent the user as a mixture of topics, i.e shared interests. Their system recommended friends for 4 million users with high recall, out performing graph-based models.

Lee et al. (2011) research presents a technique for item recommendation within social networks that matches user and group interests over time. They use an adapted LDA algorithm where two LDA models represent users and items as mixtures of latent topics over time. The time aspect is modelled in a one LDA model with timestamps and tags in timestamped items. They applied their concept to Flickr, a photo sharing social media. Their results show that mean precision above 70% for their model taking time into account for the recommendations. The approach was compared to the similar recommendation technique without the time element which still had a precision about 60%.

Chapter 3

Methodology

This chapter will present the methodology which the research project is based on. The research project followed the Design Science methodology and guidelines, and the development of the Semantic Tourist followed the system development framework presented by (Nunamaker Jr et al., 1990)

3.1 Design science

Research in the information systems discipline is concerned with people, organizations and technology. Researchers try to understand problems related to developing and successfully implementing information systems in organizations. According to Hevner et al. (2004) there are two paradigms that characterize the field, behavioural science and design science. The first has its roots in natural science research and seeks to develop or justify theories that explain or predict human or organizational behaviour. The latter, which is to be used in this thesis, is rooted in engineering and sciences of the artificial (Hevner et al., 2004, (Simo 1996)). It seeks to create innovations that effectively and efficiently solve problems for humans and organizations.

The fundamental questions in design science are ”What utility does the new artefact provide?” and ”What demonstrates that utility?” (Hevner et al., 2004). They are extracted from the fundamental principle of design science; that “knowledge and understanding of a design problem and its solution are acquired in the building and application of an artefact ”(Hevner et al., 2004). To address this, evidence is needed and Hevner et al. (2004) propose guidelines to produce this evidence.

Design science was chosen as the main research method for this project because the nature of the method goes well with the project research goals. The development and evaluation of the artefact gave increased insight for further development of the artefact. Without an artefact it would have been difficult to test the theory with users, which was necessary for theory confirmation.

3.1.1 Design Guidelines

The design guidelines presented by Hevner et al. (2004) are intended for researchers and others to understand the requirements in design science research. The following will include a presentation of the seven guidelines, an explanation of them and an assessment of how the project followed them. The seven guidelines are: Design as an artefact, Problem relevance, Design evaluation, Research contribution, Research rigour, Design as a search process and Communication of research.

Design as an artefact

The first guideline states that "Design Science research must produce a viable artefact in the form of a construct, a model, a method, or and instantiation"(Hevner et al., 2004). In their definition Hevner et al. (2004) puts less emphasis on people and the organization related to the artefact and more emphasis on the construct, model, method and instantiation. This is to address both the process of design and the designed product. An artefact is created to address a problem. In innovation, the constructs, process and models used contribute to the creation of the artefact. The process may also demonstrate that some things can be done in a more efficient way. Even if the instantiation is not suited for implementation in a real life organization, it can be an important part of the research as a "proof by construction", and is a step towards deployment.

The artefact produced in this research is a prototype of a personalized tourist application. The main research part was on the creation of the two user models, the topic model and the RDF-model. The models are representations of the user's interests and preferences elicited from the user's social network presence. They form the basis of the personalized recommendations of points of interest (POI). The application is presented in chapter 4 and the continuous evaluation of the central parts of application is described in more depth in Section 5.1 . The combination of source for the user models and the technologies used to created the models is the innovative part of the project. The implementation is described more in section 4.4. The artefact is a new take on solving the "old" problem of giving personalized content-based recommendations.

Problem relevance

The second guideline states that "the objective of design-science is to develop technology-based solutions to important and relevant business problems"(Hevner et al., 2004). In design science this goal is addressed through the construction of innovative artefacts that changes the way a problem is solved. A problem is defined as "the differences between a goal state and the current state of the system"(Hevner et al., 2004). The design-science research also has to be relevant with respect to the constituent community. The people in the community are the ones that are involved in developing and implementing the project (planners, managers, designers, implementers, operators and evaluation people).

The general domain of this project is personalization of recommender systems. The artefact created in this project tries to combine technology and sources in a new way. Furthermore the research problems that the project tries to solve are: "Using semantic technologies and information collected from the users social network profile, is it possible to generate a reliable model of a user's interests?","Does Facebook provide sufficient information to do this?" and "To what degree can the resulting model be used to generate recommendations in the tourist domain?" The main semantic technologies utilized are topic modelling and RDF-modelling. The technologies are used to elicit information from the user's Facebook profile creating two models which represent the user's interests. The models are the basis for recommending locations and events which relates to the user's interests. More detailed description of the research problem is found in Section 1.1.

Design evaluation

The third guideline proposes that "the utility, quality, and efficacy of the design artefact must be rigorously demonstrated via well-executed evaluation methods"(Hevner et al., 2004). Evaluation of the IT artefact has to establish appropriate metrics, and gather and analyse appropriate data. The artefact can be evaluated in terms of functionality, completeness, consistency, accuracy, performance, reliability, usability, fit with the organization and other relevant quality attributes.

The methods used to evaluate the artefact are described in chapter 5. The functionality, accuracy and performance is demonstrated through the user testing that was conducted. The application is a prototype and there are some features that have been left out because of time and scope restrictions. Additionally, the application's usability was not an important part of the project. Thus the usability of the application is only discussed briefly. The evaluation of completeness, consistency and usability can be found in Section 5.3.

Research contribution

The fourth guideline suggests that “effective design-science research must provide clear and verifiable contributions in the areas of the design artefact, design foundation, and/or design methodologies” (Hevner et al., 2004). It concerns how the research contributes to the knowledge base. Following this guideline you try to answer the question “What are the new and interesting contributions?” (Hevner et al., 2004). You can contribute through novelty, generality or significance of the designed artefact.

In this case it is the artefact itself and the results from the evaluation that contributes to the research. The novelty in the research is taking technologies that have been tried and proven in other domains, and combining these to create a new artefact. The most important contribution will hopefully be the demonstration that a user model can be built from ”general” information and utilized to say something about a person’s interests in the tourist domain. The research produced increases insight in the usage of topic models and how they can be used to create user models for content-based recommender systems. For more details about the application see Chapter 4.

Research rigour

The fifth guideline says that “Design Science research relies upon the application of rigorous methods in both the construction and evaluation of the design artefacts”(Hevner et al., 2004). It is about the way the research is conducted, and about the researchers skills in selecting appropriate techniques to develop or construct a theory or an artefact. He or she also has to select appropriate means to evaluate the artefact or justify the theory. Design science is mostly concerned with the machine part of the human computer interaction which is key to the success of the artefact.

The development of the artefact was iterative and incremental. This provides continuous evaluation and feedback on design and functionality, which is important because it facilitates error correction. It also makes it possible to react to changes in requirements. The product is complete when it solves the problem it was meant to solve.

The construction of the artefact is based on personalization techniques that have shown good results in other research. The topic modelling approach is presented in Section 2.5. Other user modelling approaches are described in Section 2.3 while various recommender system approaches are presented in Section 2.4. This provides a good foundation for the project. See the discussion (Section 6.2) about how well the topic modelling is fit to do this job. The use of rigorous methods for evaluation is discussed in Chapter 5.

Design as a search process

The sixth guideline tells us that "the search for an effective artefact requires utilizing available means to reach desired ends while satisfying laws in the problem environment"(Hevner et al., 2004). The searching refers to the iterative and incremental development of the artefact which facilitates continuous improvement. While the means are the actions and resources used to construct the artefact. Goals and constraints represent the ends, while laws represent the environment and everything unforeseeable.

Design-science research often simplifies the problem by dividing it into sub problems. This does not always create a satisfactory solution to the problem as a whole, but can make a good starting point. Through the search the prototype evolves from the starting point through to a satisfactory solution and on to better solutions. Researchers can use heuristics to evaluate the quality of a solution.

The project has been developed in an iterative manner (see Section 3.2.1 for the theory and Section 5.1 for the actual process). The development of the topic model was continually evaluated in order to find the parameters that gave the best results. When the model showed unwanted results on one dataset the parameters were tweaked before it was tested again. More on this in Section 5.1

Communication of research

The seventh guideline advise that "design science research must be presented effectively to both technology-oriented as well as management-oriented audiences"(Hevner et al., 2004). When presenting for a technology oriented audience it has to be in such a way that they understand the research and how to draw knowledge from it. In contrast, when presenting for management, the research should be presented in such a way that they understand how it can be used to give advantages to the organization.

For this project I have to consider two different audiences. The main focus will be to target the presentation toward the technical/research audience within the information science field. I will also have to consider presenting the concepts to the test subjects.

I have to present the research in a way that is understood, show what knowledge can be drawn from it, and suggest some guidelines or a list of advantages and disadvantages that can be used by other developers/designers. Since there is no organisation I will have to explain how it will give advantages to users, and advantages over other approaches in the field. The thesis in itself will be a technical document describing the development process and the guidelines will contribute to explain pros and cons of the artefact.

To summarize the section; what utility does the artefact provide? The artefact provides utility by performing the tasks intended and contributing to extending the knowledge-base in the different domains it touches. What demonstrates that utility? The tasks are evaluated according to "good" measures and the thesis will demonstrate it contributes to the knowledge-base.

3.2 System development

3.2.1 Theory

The Design Science methodology puts great emphasis on research rigour (Hevner et al., 2004). This project has followed the system development framework presented by Nunamaker Jr et al. (1990). The framework was chosen because it puts great emphasis on the importance of doing iterations with continuous evaluation. The following sections will describe the framework in more detail.

The first stage is to create a conceptual framework. At this stage the researcher should justify the significance of the research. It should include an investigation of the system's functionality and requirements, an understating of the system's building processes/procedures and the study of relevant disciplines for new approaches and ideas (Nunamaker Jr et al., 1990).

The second and third stage go together. The second stage is the creation of the system's architecture and the third is to analyse and design the system (Nunamaker Jr et al., 1990). At this stage the system's components should be put in the correct perspective, a specification of the functionality of the system should be created, and the relationships and the dynamic interactions between the structures should be defined. Requirements should be defined in a way such that they can be evaluated and the emphasis should be put on the new functionality.

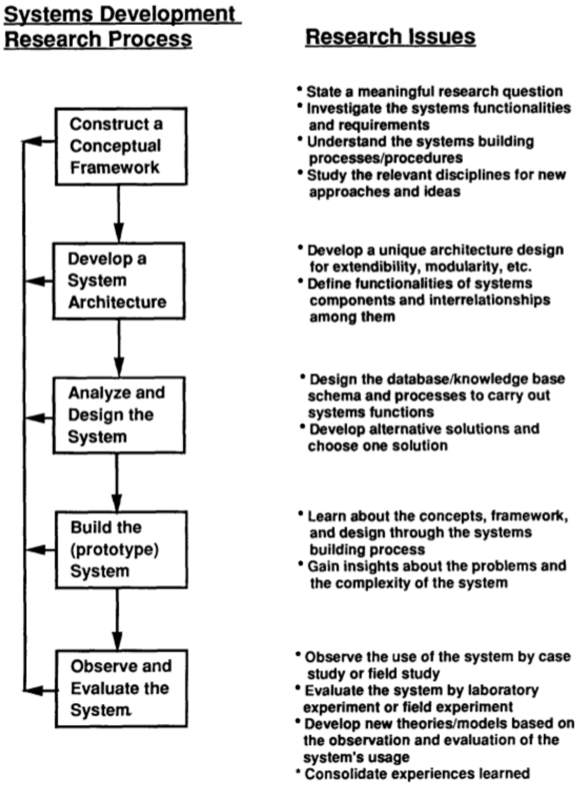

Systems Development Research Process
Research Issues
Construct a Conceptual Framework
Develop a System Architecture
Analyze and Design the System
Build the (prototype) System
Observe and Evaluate the System
• State a meaningful research question
• Investigate the systems functionalities and requirements
• Understand the systems building processes/procedures
• Study the relevant disciplines for new approaches and ideas
• Develop a unique architecture design for extendibility, modularity, etc.
• Define functionalities of systems components and interrelationships among them
• Design the database/knowledge base schema and processes to carry out systems functions
• Develop alternative solutions and choose one solution
• Learn about the concepts, framework, and design through the systems building process
• Gain insights about the problems and the complexity of the system
• Observe the use of the system by case study or field study
• Evaluate the system by laboratory experiment or field experiment
• Develop new theories/models based on the observation and evaluation of the system's usage
• Consolidate experiences learned

FIGURE 3.1: "A Research Process of Systems Development Research Methodology" (Nunamaker Jr and Chen, 1990)

The fourth stage is where the implementation happens. The implementation should demonstrate the system's design feasibility and the usability of its functions (Nunamaker Jr et al., 1990). At this stage more insight into the advantages and disadvantages of the concepts, the framework and the design is also accumulated. The experience and knowledge is helpful in the re-design of the system in the next iteration.

The last stage of the iteration is to observe and evaluate (Nunamaker Jr et al., 1990). The system is tested against the requirements for performance and usability. The development is a continuous, evolutionary process where experience from each iteration leads to further development and new ideas in the next.

3.2.2 Practises

As stated in the previous section, the first stage of the development process is to construct a conceptual framework. The investigation part of this stage was spent on reviewing articles in the relevant research fields. The review provided an overview of what techniques had been used before. Some investigation of what information is available in a user's social networks, and ways to process the information was done in this phase. The initial investigation lead to the selection of Mahout as the tool for extracting meaning from the user's information. Hence this stage also included a significant technical spike (Shore, 2007) in order to learn about topic modelling in general and Apache Mahout in particular to get an understanding of it's functionality.

A couple of use case scenarios was created in the early stages of the development process. The scenarios describe the human activities and tasks which is related to the use of the system. These scenarios allows for further exploration and discussion of context, needs and requirements. The scenario should not describe the use of software or technology (Sharp, 2007). The choice of an Android device as the client for the application was based on these scenarios. A mobile client was chosen because there is not a lot of input required for the application and a user would have it at hand when travelling .

The following is one of the scenarios: *"A business man is on a trip to Bergen. He has some spare time and starts The Semantic Tourist. It suggests a round trip in Bergen including a visit to the Hanseatic museum and going to Fløyen. With the recommendation of going to Fløyen it has given a remark about the possibility of rain. The business man rejects the suggestion and chooses another option instead. The application also suggests restaurant with traditional Norwegian food, as a last stop before going back, because it know that the business man likes to try out the local cuisine when travelling. A restaurant serving Italian food is also recommended, because the application knows that the user is from Italy."*

Continually developing the architecture and analysing the design helped in improving the results throughout the development process. Either through adding or removing input information, using fewer or more topics or topic words, trying different queries and adding other sources of points of interest. See Chapter 4 for implementation details. The building of the prototype would provide more insight, and running the prototype and evaluating the results gave confirmation or retraction of the design. See Section 5.1 for more details about how the application was tweaked in order to produce the best possible results.

Chapter 4

The semantic tourist application

This chapter will describe the implementation of the Semantic Tourist. The first section will place the project in the theoretical framework. The second section will present the architecture to give an overview of the application. After the architecture section, the data sources will be presented along with explanations for choosing the sources. Next follows a presentation of the server and the client. These sections also explains various choices made in implementation process. The final section of the chapter will present the development environment of the project.

4.1 Choosing personalization approach

The goal of the Semantic Tourist was to use semantic technologies in order to create an application that collects data from the user's social network presence and generates a model of the user's interest which is used to recommend personalized points of interests in the tourist domain. The following explains how this project fit into the three dimensions of personalization presented in Section 2.2. This project implements personalization of an *individual user* through *implicit extraction* of information from the user's Facebook profile. The recommendations are based on the information, or *content*, itself.

Traditional tourist guides and applications (apps) usually gives recommendations based on the stereotype model. Stay.com is an example of this type of tourist application. It gives the same recommendations to all users, points of interests that the average tourist would want to visit. The category system in traditional systems are also based on the tourist stereotype.

Ratings from users are a common way of recommending specific POIs. Stay.com also give their recommendations through the ratings of other user. It is however not a given that the current user has the same preferences as other users. An individual user model was chosen for this project to determine whether an individual user model could give better results than the stereotype model. To leverage the cold-start problem that exist with individual user models, this application use the users' social network to gather information about the users' interests. The prerequisite about using the users social network presence also opened for using the users friends information to generate recommendations. However this approach was rejected on the same principle as the stereotype approach. A person doesn't necessarily like the same things as other users, even if they are his or her friends.

The prerequisite of the project was to use the user's social network to elicit the required information. This is why implicit information extraction was chosen over explicit user modelling. The use case scenario and the choice of a mobile client suggests that the user should not be forced to enter a lot of information since this will decrease the usability of the application. Another argument for using the user's Facebook information was that the system should not be dependent on explicit user input.

One advantage of using Facebook as a source is that the content is to some degree categorised. The RDF-model approach in this project leverage this feature. See more in Section 4.4.2

The advantage of the topic modelling approach is that the system can extract semantic meaning from the unstructured descriptions of the user's interests. No prior annotation or labelling of documents are required. This way the application can utilize the Facebook information that is not structured. The topic modelling can determine what type of business a company is, or the theme of a TV show, and then generate a list of words related to the topic even if the Facebook-category or the title of the page doesn't say anything about it. The advantage of using the Latent Dirichlet Allocation topic modelling algorithm over other topic modelling algorithms is that it generates topics across the document collection, not just in single documents Blei (2011). Different drawbacks to using this method are discussed in Section 5.1.

Other content-based recommender systems usually relies on the user's previous behaviour from the same domain. A problem with this approach is that it recommends only the what it knows the user likes, no "new" items are recommended. This system however, generates recommendations based on preferences from a different domain. A more general domain. An advantage of this method is that the system might recommend things the user would not normally think of in the tourist domain, because it knows what the user likes, in this other, more general domain. An example from the test; one user who enjoys playing the piano got recommended piano bars, which he explicitly indicated was a good, surprising recommendation.

4.2 Architecture

The system architecture is built on the Model-View-Controller[1] pattern. The Spring framework was chosen for easy set up of the Java Web server. Spring Framework contains a Spring implementation of the software architecture pattern Model-View-Controller (MVC). It was used to implement a service endpoint for the application. The view was implemented on the Android client and the controller(s) and model on the server. There are several controller, classes on the server which handles different functionality. One controller handles the storage of the Facebook user information, one controls the topic model, one controls the queries against LinkedGeoData and one manages the queries against eventful.com. In keeping the responsibilities separate, the architecture facilitates easy integration of new input sources e.g. Twitter[2] and Tumblr[3], and new third party data sources e.g. sindice.com[4] and meetup.com[5]. It also facilitated the relatively easy change of topic modelling library during development (see more in Section 4.4.1).

Spring framework facilitates communication between server and client, allowing easy set up through annotations-based rendering of URL's to the right controller. The Android client sends HTTP requests containing information serialized in JSON. JSON stands for JavaScript Object Notation and is an open standard widely used to send data over the HTTP protocol (Little et al., 2010). REpresentational State Transfer (REST) is an architecture approach to Web-based applications (Little et al., 2010) which sends information through HTTP, usually using JSON or XML[6]. In a ”REST like” manner, the controllers are implemented to receive data via HTTP POST on URLs, e.g “/server/me_likes”. With this interface the server could receive request from any third party client as long as it forms the request according to how the implemented protocol.

Mallet is used to create the topic model which is the basis for locating points of interest (POI) the application recommends. Mallet or ”MAchine Learning for LanguagE Toolkit” is an open

[1] MVC explained: http://www.oracle.com/technetwork/articles/javase/index-142890.html
[2] https://twitter.com/
[3] https://www.tumblr.com/
[4] http://www.sindice.com/
[5] http://www.meetup.com/
[6] http://en.wikipedia.org/wiki/Xml

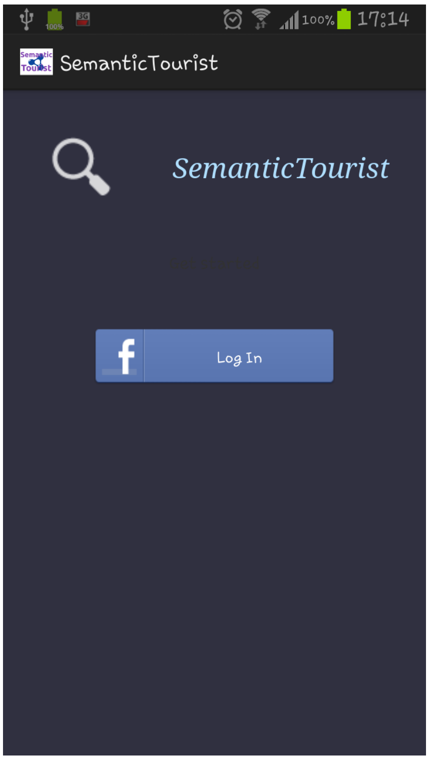
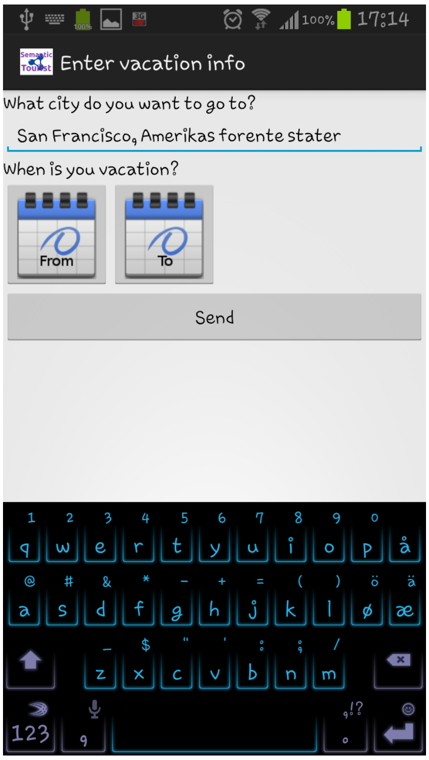


FIGURE 4.1: Screenshot of login fragment



What city do you want to go to?

San Francisco, Amerikas forente stater

When is you vacation?

FIGURE 4.2: Screenshot of "enter vacation" fragment

source Java-based package for statistical Natural Language Processing (McCallum, 2002). Document classification, clustering, information extraction and topic modelling are among machine learning features of Mallet. When the server receives the user's information from the client, the information is filtered and made into a the document collection Mallet uses as input. The topic model resulting from running the topic modelling algorithm is transformed into a set of query terms used to query the LinkedGeoData SPARQL endpoint.

Apache Jena is the Java framework used for building the RDF-model which is the basis for locating the event POIs the application recommends. Jena provides a collection of tools and libraries which is used to develop Semantic Web and Linked Data applications, servers and tools (The Apache Software Foundation, 2013b). This project uses Jena to create and store an RDF-model of the user's Facebook information. The titles of the user's music and activity interest, e.g names of bands, artists etc. and activity names like basketball, snowboarding, programming etc, are selected from the RDF-model and used as search term in retrieving events from the eventful.com API. The application also uses the Jena SPARQL query service libraries for querying the LinkedGeoData SPARQL endpoint over HTTP.

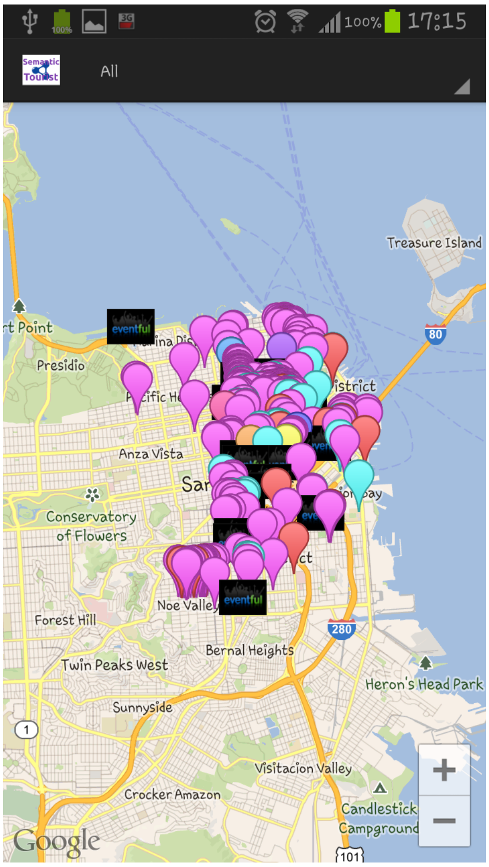


Figure 4.3: Screenshot of "Enter vacation" fragment

The view of the application is the Android client. The first screen displayed when the application starts is the log in fragment(see Figure 4.1). A fragment represents a part of the user interface in an activity (Android Developers, 2013b). An activity is a component of the application which represents a single screen (Android Developers, 2013a). An example activity is the login activity that starts the application. The user logs in with his or her Facebook account information and is taken to the next activity where the information about where and when the user is going on vacation is given as input (see Figure 4.2). When the user has logged in an asyncTask starts to check if the user's information has been updated today. If it is not updated or it is a new user, it retrieves the user's Facebook information and sends it to the server. When the user's information is updated on the server and the vacation destination is entered the user clicks send. A map of the vacation location is then displayed. Android AsyncTasks [7] requests the information about the location and event POIs. An AsyncTask performs background operations, like requesting information, and publish the results to the user interface. The results are presented as markers on the map as shown in Figure 4.3.

4.3 Data sources

4.3.1 Facebook

"Facebook helps you connect and share with the people in your life" (Facebook.com, 2013). - Is what Facebook says about themselves on the front page. Facebook has about 665 million daily active users, 1.11 billion monthly active users (Facebook, 2013b) and on average 4.5 billion likes generated daily (Facebook, 2013d). In the profile users can express their preferences through listings of favourite Movies, Books, TV shows, Artists etc. They can also list their interest and

[7] http://developer.android.com/reference/android/os/AsyncTask.html

activities that they like. Facebook also has opportunity for sharing status updates, checkins for places the user is visiting and sharing photo etc.

When choosing what social network should be the main source of information it was important that the social network has a large user group and a profile that allows for expressing the user's interests and preferences. Personal blogs and Twitter were among the other social networks that were considered. Even if there are many users in these social networks, it is not as widespread as Facebook. Personal blogs and Twitter also do not have the same diverse options for explicit sharing of interests, as Facebook has. Hence Facebook was chosen because of its' large user group and rich user profile.

Accessing the user's information is fairly easy through the one of Facebook's System Development Kits (SDK). They offer easy access to the Facebook Graph API through "easy-to-use" methods (Facebook, 2013a). The Graph API is a HTTP-based API which gives access to e.g the user's profile information (Facebook, 2013e). The Facebook Explorer allow the developer to explore what information is available with different user permissions. Section 4.5.1 describes how it is implemented in the Semantic Tourist.

Manago et al. (2008) claim that people present an idealized version of themselves in online social networks. While others ((Vazire and Gosling, 2004),(Back et al., 2010) (Tosun, 2012)) claim that that people express their real personalities. This is something to be aware of in the evaluation of the application. However, this study will not discuss this any further because it is outside the scope of the research.

4.3.2 LinkedGeoData (LGD)

LinkedGeoData is a Semantic Web project that "uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles"(LinkedGeoData.org, 2013). OpenStreetMap is a project that creates and distributes free geographical data for the world (OpenStreetMap.org, 2013). The LinkedGeoData project adds a spatial dimension to the Web of data and links to other knowledge bases.

LinkedGeoData was chosen because it contains all the geographical data from the OpenStreetMap, including names of shops and restaurants. Compared with other tourist APIs it is not restricted to the obvious tourist locations. It was also an appropriate choice because it is open linked data which is ready to be linked with other semantic data. By choosing LGD as main source it was possible to use other sources (e.g. DBpedia) for linking extra information about the locations.

4.3.3 DBPedia

"DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make it available on the Web"(Dbpedia.org, 2008). The information from Wikipedia is semantically lifted and transformed to RDF which is made available under a Creative Commons license. It is accessible online through several interfaces including a SPARQL endpoint.

DBpedia was chosen as a secondary source to LinkedGeoData because it has a lot of information about buildings, attractions etc. LGD also interlinked with DBpedia pages.

4.3.4 Eventful.com

Eventful.com[8] is a website where it is possible to search for events based on location, date, keyword and category. The Semantic Tourist uses the eventful.com REST API to retrieve events that are related to the music, sport and interest related likes of the user.

Eventful.com was chosen because it contains event information which would give the Semantic Tourist application a feature traditional tourist applications and guidebooks do not have. If one of the bands I like play in the city I'm going to, it would be fun to know about.

On the eventful.com web page it is possible to search for events browsed by category. The response from the REST API however, does not contain information about which category the event belongs to. There was unfortunately not time to implement the eventful.com category system in an ontology which would most likely would have made the search for events more accurate.

There are various other event APIs like meetup.com[9] which, if there was time, would have been added as data sources. The drawback with these APIs is that the data is not semantically annotated. However the data they provide can be transformed to RDF.

[8] http://eventful.com/

[9] http://www.meetup.com/

4.4 Server

4.4.1 Topic model

This section will describe the different implementations of the topic model that were tested during development of the project. The goal of running the topic modelling was to create a set of words that could be used to retrieve points of interests (POI) from third party sources. Specifically, the generated words should help retrieve nodes in the LinkedGeoData source. Examples of topic model output and the related evaluations is presented in Section 5.1.

Natural Language Processing with a topic modelling algorithm was chosen as the main method to elicit semantic data from the user's Facebook profile. The decision was based on the fact that most of the user information from Facebook is unstructured which means that the computer has little initial information to work with. An advantage with using the Latent Dirichlet Allocation algorithm over other topic model algorithms, is that "all the documents in the collection share the same topics"(Blei, 2011). This way if there is a topic cutting across documents, the algorithm should be able to find this connection.

The topic model algorithm extracts words from the documents that represent the topic of the text. They describe what the document is about. These words are then used by the computer to query for points of interests that are related to the contents of the document. Because of the limited time and scope of the project it was important that the application was built on top of already existing libraries that were open source and free. This was especially important in choosing the topic modelling tool, because it is the most advanced and would require the most time to write from scratch. By using an already existing tool the project could focus on making the different technologies work together in the best way possible.

Apache Mahout and Mallet were two of the already existing libraries that met the necessary criteria of having a command line tool to allow for easy testing of the library, and being integrable with the back-end of The Semantic Tourist. Mahout was chosen as the starting tool because of it's apparently superior documentation and because it has a large user base. Hence seemed easier to get started with. Mallet was implemented instead at a later stage (see Section 4.4.1.3 and Section 5.1).

Apache Mahout is a machine learning library which implements various learning algorithms like collaborative filtering, K-means clustering and Latent Dirichlet Allocation (LDA) to name a few

Included attributes	name, description, about, Website, link, category, personal info, general info, affiliation, mission, products, company overview

Table 4.1: List of Facebook page attributes included in the document collection used as input to the topic modelling algorithm.

(Mahout, 2011). The LDA implementation is the one used in this project. To run the algorithm, Mahout needs the input documents to be indexed. This project used Solr, as recommended by the Mahout homepage, to generate this index. Solr is an open source search platform from the Apache Lucene project. The project uses the ”near real time indexing” and the REST-like API to post JSON over HTTP directly from the client to the server (Foundation, 2012). This feature facilitates the indexing of documents. The index is then used to create the topic model.

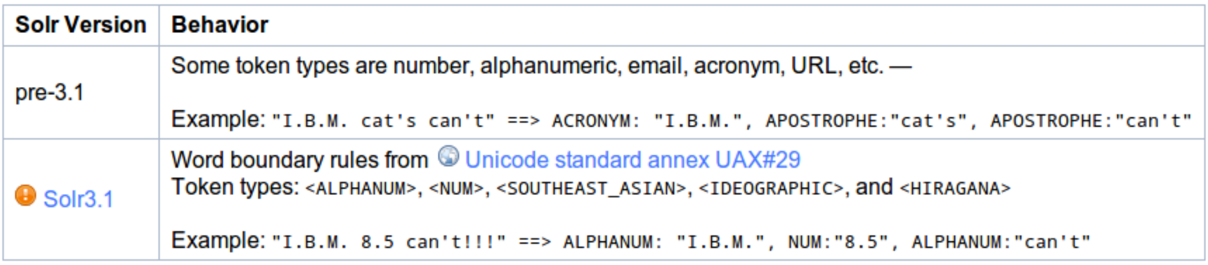
4.4.1.1 Preprocessing documents before topic modelling

The JSON returned from the Facebook Graph API contains the different attributes that together make up a Facebook page. These are just some of the attributes: name, description, number of users who like the page, about, category, opening hours (if it is a business), ”talking about count” etc. Furthermore all the pages can contain different attributes. Hence making sure that only information describing the page was indexed, was essential to avoid noise in the topic model. In this case noise refers to words that should not be used in the query. Only attributes describing features of the page were used as input. See Figure 4.1 for a list of the attributes.

Solr was configured to tokenize the text and remove stop words before the indexing. Without the tokenizer filter, acronyms and words with apostrophes would be split into separate tokens. However with the filter these are left untouched to give more meaningful tokens. See Figure 4.4 for an example. The stop word filter removes commonly used words that do not provide any meaning by itself in the text. Examples can be found in Figure 4.5

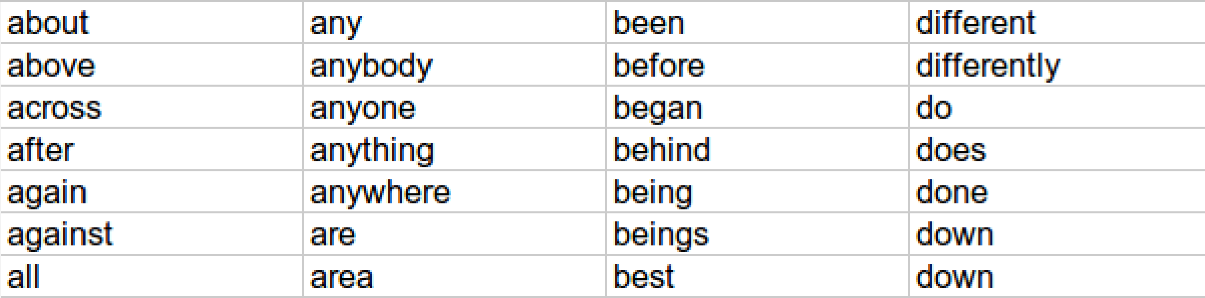
4.4.1.2 Topic model adjustments

The index produced by Solr is passed to Mahout in order to generate the topic model. Choosing the right number of topics to be generated is more art than science (Hall and Kanjilal, 2013). Because the topic modelling in this project is intended to create terms which later can be used in queries against other data sources, mainly LinkedGeoData, the input to the algorithm also



Solr Version	Behavior
pre-3.1	Some token types are number, alphanumeric, email, acronym, URL, etc. — Example: `"I.B.M. cat's can't" ==> ACRONYM: "I.B.M.", APOSTROPHE:"cat's", APOSTROPHE:"can't"`
Solr3.1	Word boundary rules from Unicode standard annex UAX#29 Token types: `<ALPHANUM>`, `<NUM>`, `<SOUTHEAST_ASIAN>`, `<IDEOGRAPHIC>`, and `<HIRAGANA>` Example: `"I.B.M. 8.5 can't!!!" ==> ALPHANUM: "I.B.M.", NUM:"8.5", ALPHANUM:"can't"`

Figure 4.4: Example of how the Solr tokenizer works



about	any	been	different
above	anybody	before	differently
across	anyone	began	do
after	anything	behind	does
again	anywhere	being	done
against	are	beings	down
all	area	best	down

Figure 4.5: Examples from the stop words list used by the topic model algorithm.

Type of change	**What changed**
Topics filter	x number of topics, number of documents / x
Words filter	top 2 highest probability, topics with higher than x probability, x number of words
Documents	All likes, all likes except x categories, separate categories, Web sites of likes (about pages, paragraphs and titles), place pages of checkins, statuses
Categories	Restaurants and Bars, Interests, Interests with Web pages, Local business, Interest and Local business, All except interests, No media, No artist with Web pages, All, All with Web pages

Table 4.2: This table show the various changes that were tried when implementing the topic model

plays an important part in creating the topics. Table 4.2 show a summary of the different things that were tried.

The first row in Table 4.2 show the different ways that were tried to decide how many topics the algorithm should generate. It ranged from 3 to 40 topics. For each number of topics, the algorithm was also tested with from 2 to 10 words for each topic. The probability filter was rejected because the variation in the probability of a word belonging to a topic varied too much from topic to topic. The third and fourth row show the different combinations of input documents tested. Documents refer to the different information collected from Facebook, and the categories are the ones defined by Facebook. The final test of the Semantic Tourist was run with the following parameters: number of topics = number of input documents divided by

Movie, Book, Author, Movie genre, Artist, Band, Tv show, Musical genre, Musician/band, Concert venue, Movie Character, Writer, Teacher, Producer, Actor/director, App Page, Community

Table 4.3: Set of Facebook categories that was not included in document collection used as input to the topic model algorithm.

5, number of words for each topic = 3, input documents = all likes except the ones in table 4.3, pages of places the user checked in at and the 25 last status updates. The number of status updates was limited to 25 to limit the time it would take the server to analyze all the information running on my laptop. In a deployment with more cpu power available, this number would preferably be much higher, and possible include all status updates.

4.4.1.3 Changing topic model library

After some testing Mahout still was not performing satisfactory. Most of the “noise words” were gone, but there were still some words that did not fit with the topics. So I was advised to try Mallet instead, to see if this would solve the problem. The system was designed so that changing the library used to generate the topic model was a simple matter of exchanging the class running the Mahout library with one running Mallet. The new library had a different set-up from Mahout, thus some iterations of configuration and testing was necessary. After the change, the topic modelling analysis ran noticeably faster and the “noise words” were not a part of the topic model.

4.4.1.4 Pre-query topic words filter

The words from the topic model are used as base for the query terms in the SPARQL queries against LinkedGeoData (LGD). Before the terms can be used in the queries, the topic words are run though another filter. This filter removes any words with two or less characters because they do not have any meaning by themselves. The filter also adds the singular form of any noun to the list of query terms. This is because words in the topic model originates from unstructured text where nouns will often be in their plural form. However the LGD categories are modelled with the singular form of nouns, so the filter adds the singular form in order to match the categories. The LinkedGeoData is written in English, hence all the topic words also had to be translated into English to match the resources before the query was executed. See more evaluation details in Section 5.1.2

4.4.1.5 LGD queries

The server runs two queries against LinkedGeoData, one where the topic words (called query term when its used in the query) should match the LGD-category (also refereed to as "category-query") and one where it should match the label/title of the resource (also referred to as "label-query"). Both queries takes three parameters; a query term, and longitude and latitude for the location the user wants to visit. It selects nodes where the geo-coordinates of the node are within a set radius of the geo-coordinates provided as input to the query. Figure 4.1 shows the "label-query" written in turtle format. The query is created such that if the query term is "sport", then the result returned will contain all nodes where sport is the category, or sport is one part of the category, eg. SportCenter. It will however not match "HairShop" if the query term is "air". It uses a regular expression (regex)[10] which matches "whole words only". The label query selects all nodes where one of the words in the label matches the query term, eg. "food" matches "Spicy food plus". The label literals are often written in the language of the originating country. Therefore the label query terms are translated with Google Translate API, to match the language of the label. LinkedGeoData mostly have rdfs:label in the local language despite rdfs:label facilitating labels in multiple languages.

```
select * from <http://linkedgeodata.org>
{
  ?node rdfs:label ?label .
  ?node rdf:type ?type .
  ?node lgdo:directType ?directType .
  ?node geo:lat ?latitude .
  ?node geo:long ?longitude .
  ?node geo:geometry ?g .
  Filter(<bif:st_intersects> (?g, <bif:st_point> ("+longitude+", "+latitude+"), " + radius + ") ).
  Filter( regex(?label, "\\b"+topic+"\\b, "i")) .
  OPTIONAL {?node lgdp:cuisine ?cuisine }.
  OPTIONAL {?node lgdp:Website ?Website }.
  OPTIONAL {?node lgdp:tel ?tel }.
  OPTIONAL {?node lgdp:phone ?phone}.
  OPTIONAL {?node owl:sameAs ?dbpedia}
}
```

Listing 4.1: Java code for the SPARQL select query used in the Semantic Tourist to select label directType and geographical information for the nodes which are in the specified radius of the requested location

[10] http://www.regular-expressions.info/wordboundaries.html

4.4.2 RDF model

```
@prefix fb: <http://graph.facebook.com/> .
@prefix st: <http://master.lisahalvorsen.com/semantictourist/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

fb:123456 st:likes fb:103818329656399
fb:103818329656399 foaf:name "Android"
                st:category st:Interest

fb:123456 st:likes fb:106152776081709
fb:106152776081709 foaf:name "The Pogues"
                st:category st:Band

fb:123456 st:likes fb:111528558865423
fb:111528558865423 foaf:name "Jokke"
                st:category st:Musician_band

st:Band rdfs:subClassOf st:music
st:Artist rdfs:subClassOf st:music
st:Musical_genre rdfs:subClassOf st:music
st:Musician_band rdfs:subClassOf st:music

st:Interest rdfs:subClassOf st:interest
st:Sport rdfs:subClassOf st:interest
```

LISTING 4.2: Example RDF model. The user with id 123456 likes Android The Pogues and JokkeȦndroid has the category “st:Interest” which is a rdfs:subClassOf interest. The Pogues has category “Band” and which is a rdfs:subClassOf “st:music”̇“Jokke” has category “st:Musician_band” which also is a rdfs:subClassOf “st:music”

The Semantic Tourist creates two types of points of interest. One is the kind presented above, and the other is events related to the users music and activity interests. These event points of interest are generated from the other user model in the Semantic Tourist, the RDF-model. This is a simple RDF-model of the user’s Facebook likes. The JSON list of likes retrieved from Facebook is modelled in RDF.

The user’s Facebook id is used as resource identifier for the user, and the likes are added with foaf:name and category property as shown in the example in Listing 4.2. Friend of a Friend (FOAF) is a Semantic Web project for creating machine readable Web pages containing descriptions of people, links between people and things people create and do (project). The model is stored in an Apache Jena TDB store [11]. The Facebook-categories related to music

[11] http://jena.apache.org/documentation/tdb/index.html

and interests are modelled as subclasses of the Semantic Tourist classes music and interest(see Figure 4.4). Hence the model contains rdfs:subClassOf relationships between the Semantic Tourist (st) classes and the Facebook category classes. The server use a Jena SPARQL query against the model to select the name of all nodes which are a subclass of st:music or st:interest. The name of the node is then used as query term in an eventful.com REST API request. The request also has options for geographical location and time period for the vacation.

Class	**music**	**interest**
subclass	Artist Musical genre Musical band Band	Interest Sport

Table 4.4: Shows the class hierarchy in the RDF model of the Semantic Tourist. It is used to recommend events based on the user's likes from these Facebook categories.

4.5 Client

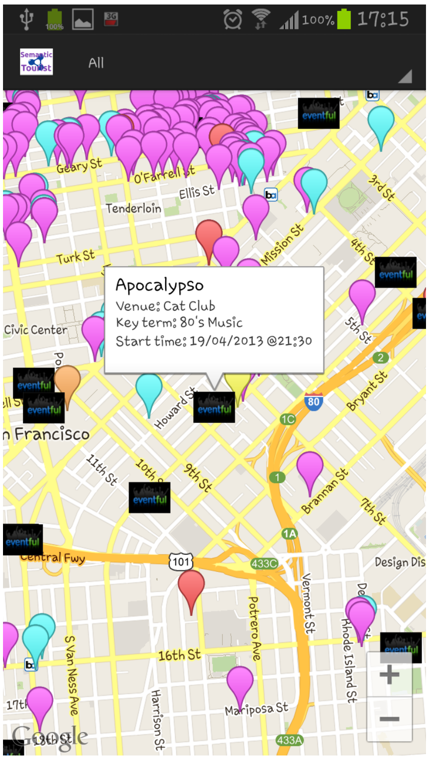
4.5.1 Android

The use case for the Semantic Tourist is in planning your vacation, at home or in the hotel room. The choice of client interface was based on the use cases and the nature of the application. Because the Semantic Tourist is about finding location and discovering what is in a given location the client was built for a mobile device. For this type of application a mobile client is suitable because it does not demand a lot of typing to use the application, and you do not have to bring the computer.

There are frameworks for working with semantic data on the Android device, but the topic modelling should be run on a machine with more capacity, so it was decided to let the client handle the communication with Facebook and let the server handle modelling and retrieving of points of interests.

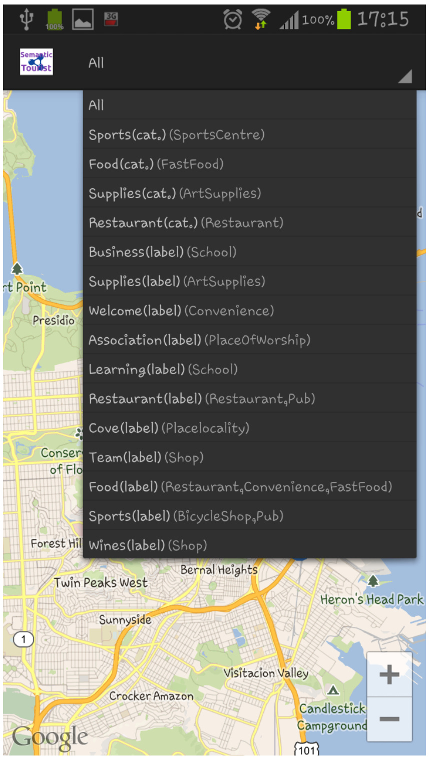
Facebook Android SDK

For easier integration against Facebook the project uses the Facebook SDK for Android. It facilitates login and requests against the Facebook Graph API. The login functionality handles the authorization against Facebook, what user information is accessible to the application and manages the Facebook user session. When the user logs in he or she is asked to give the application permission to access likes, statuses and checkins from the Graph API. The request



Apocalypso
Venue: Cat Club
Key term: 80's Music
Start time: 19/04/2013 @21:30

FIGURE 4.6: Screenshot of map with topic model markers (coloured) and events (black).



17:15
All
All
Sports(cat.) (SportsCentre)
Food(cat.) (FastFood)
Supplies(cat.) (ArtSupplies)
Restaurant(cat.) (Restaurant)
Business(label) (School)
Supplies(label) (ArtSupplies)
Welcome(label) (Convenience)
Association(label) (PlaceOfWorship)
Learning(label) (School)
Restaurant(label) (Restaurant,Pub)
Cove(label) (Placelocality)
Team(label) (Shop)
Food(label) (Restaurant,Convenience,FastFood)
Sports(label) (BicycleShop,Pub)
Wines(label) (Shop)
Presidio
Forest Hil
Twin Peaks West
Bernal Heights
Heron's Head Park
Sunnyside
Visitacion Valley
Crocker Amazon
Candlestick Campground
Google

FIGURE 4.7: Screenshot displaying example dropdown-list of categories.

for this data is facilitated by batch request provided by the SDK. The information retrieved from Facebook is sent to the server through HTTP POST's.

During the development phase Facebook had two upgrades. In November they published version 3.0 Beta 2 of the SDK and in December they released the 3.0 final version (Facebook, 2013f). The 3.0 version had new ways of sending requests through callbacks and listeners, and new interfaces for handling reading and writhing of Facebook data. The new versions forced upgrades and refactorization on the Semantic Tourist too.

Google GeoCode

When the user has successfully logged in, the screen from Figure 4.2 is shown. The application uses the internal sensors of the phone to access the current location of the user. This is to make easy for the user if he or she wants to use the application to search for POIs near their current location. When the user has entered a location the application uses Google GeoCoding API to find the exact coordinates of the location the user wants to visit. A request is then sent to the Semantic Tourist Server to retrieve the POIs for this user.

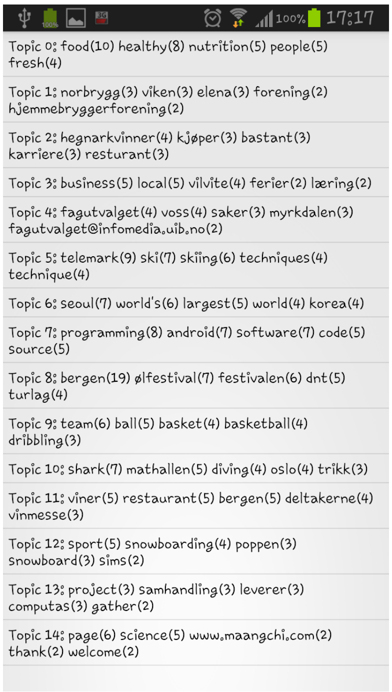
Google Android Maps API

The recommended points of interest (POI) are presented as click-able markers on a Google Map. The coloured markers are location POIs from LinkedGeoData, and the black ones are events from eventful.com. The information windows attached to the markers contains additional information from LGD. If the POI has to a related Wikipedia article, the URL is collected from the triple interlinking LGD and DBpedia, and the Wiki-article is displayed in browser. The event POIs are also click-able and the page of the event form eventful.com is displayed in the browser.

It is possible to navigate between categories using the action bar dropdown-list. The list is sorted based on the type of query. The "category-queries" are displayed on top. The reason for this was an assumption that these POIs would more accurate than the label-query. The topic word used in the query against LGD is used as the category name. The Liked Geo Data category and the type of query is also displayed in the list. Figure 4.6 and Figure 4.7 show screenshots from the map and dropdown-list in the Android client.

Topic model view for testing

A view containing the user's topic model was created for testing purposes. This view allowed the test subjects to look at the topic model generated based on their Facebook profile. This was done in order to let the user evaluated the topic model. While looking at the model the users could also find the topic related to each category in the application.



Topic 0: food(10) healthy(8) nutrition(5) people(5) fresh(4)

Topic 1: norbrygg(3) viken(3) elena(3) forening(2) hjemmebryggerforening(2)

Topic 2: hegnarkvinner(4) kjøper(3) bastant(3) karriere(3) resturant(3)

Topic 3: business(5) local(5) vilvite(4) ferier(2) læring(2)

Topic 4: fagutvalget(4) voss(4) saker(3) myrkdalen(3) fagutvalget@infomedia.uib.no(2)

Topic 5: telemark(9) ski(7) skiing(6) techniques(4) technique(4)

Topic 6: seoul(7) world's(6) largest(5) world(4) korea(4)

Topic 7: programming(8) android(7) software(7) code(5) source(5)

Topic 8: bergen(19) ølfestival(7) festivalen(6) dnt(5) turlag(4)

Topic 9: team(6) ball(5) basket(4) basketball(4) dribbling(3)

Topic 10: shark(7) mathallen(5) diving(4) oslo(4) trikk(3)

Topic 11: viner(5) restaurant(5) bergen(5) deltakerne(4) vinmesse(3)

Topic 12: sport(5) snowboarding(4) poppen(3) snowboard(3) sims(2)

Topic 13: project(3) samhandling(3) leverer(3) computas(3) gather(2)

Topic 14: page(6) science(5) www.maangchi.com(2) thank(2) welcome(2)

FIGURE 4.8: Screenshot of example topic model displayed to the user in the test session.

4.6 Development environment

Ubuntu

The development of this project started on the Windows 7 operating system. However due to problems with running the Apache Mahout topic modelling algorithm on Windows I had to switch to a Linux distribution. Hence most of the project was developed on Ubuntu 12.04.

SpringSource Tool Suite

The Semantic Tourist was developed in the SpringSource Tool Suite (SpringSource.org, 2013) (STS) integrated development environment (IDE). STS is a Spring adaptation of the open source IDE Eclipse. STS has better support for dependency management through Maven and Web-server deployment to Apache Tomcat.

Apache Maven

Apache Maven is a software project management and comprehension tool (The Apache Software Foundation, 2013a). Maven was used to handle dependencies and build the project.

Apache Tomcat 7

Apache Tomcat 7[12] was used to deploy the server. STS has automatic build and deployment to Apache Tomcat which was time saving during the development.

Sublime Text 2 / Latex

This thesis was written in Latex. "LaTeX is the de facto standard for the communication and publication of scientific documents" (Lamport, 2010). Sublime Text 2[13] was used to edit the latex.

Git/Bitbucket

Both the application and the thesis used Git[14] as the version control system and Bitbucket[15] as Git repository.

[12] http://tomcat.apache.org/index.html

[13] http://www.sublimetext.com/

[14] http://git-scm.com/

[15] https://bitbucket.org/

Chapter 5

Evaluation

This chapter will present the evaluation and the evaluation process of the project. It is divided into two main sections, evaluation during development and final evaluation. The evaluation conducted during the development process was done to improve the accuracy of the application and provide the best possible data for the final evaluation. The aim of the final evaluation is to gather data that can help answer the research questions presented in section 1.2.1.

5.1 Evaluation during development

5.1.1 Test design

The evaluation during development focused on adjusting the number and type of documents going into the topic model, finding the right parameters for the topic model and creating query words that could be used as query terms against LGD's SPARQL endpoint. All evaluations was with a single goal in mind; presenting the user with the most relevant points of interest possible.

The data set (Facebook-profile) used in the evaluation during development was the same throughout the whole development process. In the start phase of the development it was not necessary to use other data sets for testing because I was only trying to adjust the topic model so it did not contain unwanted words. When I started looking at ways to exclude unwanted POIs (see section 5.1.2) having access to other Facebook profiles could have been helpful. However because this is a prototype application and because of the time and resource restrictions of

Local business or place
Company, organization or institution
Brand or Product
Artist, band, or public Figure
Entertainment
Cause or Community

Table 5.1: Main Facebook page categories (Facebook, 2013c). When creating a Facebook page one of the subcategories of these categories have to be chosen for the page.

the project, only one data set was used during development. Five data sets were tested in the pilot test, which resulted in removal of some additional information from the input to the topic modelling (see Section 5.2.3).

5.1.2 Iterations

After choosing Facebook as the information source I started looking at what information could be elicited. The favourites and likes were a natural first choice, because these are Facebook-pages about items which the user has explicitly added to his or her profile, indicating an interest in these items. A Facebook-page can be about any concept that falls into one of the six main Facebook-page categories listed in Table 5.1. The information describing the contents of these pages are unstructured, which is why topic modelling was needed to discern the semantic structure and what they were actually about.

The goal of the topic modelling was to find the input documents and parameter/filter combination that would create topics with as few unwanted words as possible. Unwanted words in this context are words that do not make sense in the topic model context. In topic 0 in Table 5.2 most of the words are about skiing, but there are a few unwanted words that do not belong to this topic; “welcome”, “recipe”, “please” and “poesi”. In topic 2 in Table 5.2 the main theme is from the TV show House, but the words; “eight” (8 seasons), “edelstein” (name of actor) and “epps” (name of actor) is harder to place in the topic, and “dykking” does not belong to this topic. The parameter/filter are the ones presented in Section 4.4.1.1 and Section 4.4.1.4. It was also important to generate words that would be useful query terms. That is, words that would be equal to LinkedGeoData categories or words in the title of the location.

The first iteration of developing the topic model was about getting as much information about the user as possible. Some of the “likes” pages did not contain much information describing what the page was about. This is the why the related web-pages of every “likes” pages that

Topic	Topic words
Topic 0	telemark [$p(telemark \mid topic_0)$ = 0.036089608981484875
	skiing [$p(skiing \mid topic_0)$ = 0.029268143573455776
	ski [$p(ski \mid topic_0)$ = 0.025602759074976972
	techniques [$p(techniques \mid topic_0)$ = 0.014435843592593948
	welcome [$p(welcome \mid topic_0)$ = 0.014263724260675924
	technique [$p(technique \mid topic_0)$ = 0.010955068809372312
	recipes [$p(recipes \mid topic_0)$ = 0.006446937359588861
	please [$p(please \mid topic_0)$ = 0.004982880292778519
	poesi [$p(poesi \mid topic_0)$ = 0.003608960897114649
	skianlegg [$p(skianlegg \mid topic_0)$ = 4.7981169634592524E-27
Topic 2	dr [$p(dr \mid topic_1)$ = 0.048642064058526106
	house [$p(house \mid topic_1)$ = 0.03858833177990589
	hospital [$p(hospital \mid topic_1)$ = 0.022234744034333554
	eight [$p(eight \mid topic_1)$ = 0.008875191736627
	drama [$p(drama \mid topic_1)$ = 0.00459962024670554
	drug [$p(drug \mid topic_1)$ = 0.004408875025622329
	edelstein [$p(edelstein \mid topic_1)$ = 0.0043195024854646635
	epps [$p(epps \mid topic_1)$ = 0.0043195024854646635
	dykking [$p(dykking \mid topic_1)$ = 4.7695184599681E-7
	eksperter [$p(eksperter \mid topic_1)$ = 3.192441244231328E-13
Topic 7	brewery [$p(brewery \mid topic_7)$ = 0.0665896660550878
	craft [$p(craft \mid topic_7)$ = 0.060045839235203466
	beer [$p(beer \mid topic_7)$ = 0.04228665967061534
	beers [$p(beers \mid topic_7)$ = 0.03322511202767292
	brewers [$p(brewers \mid topic_7)$ = 0.02421442402309501
	breweries [$p(breweries \mid topic_7)$ = 0.02421442402309501
	brewpub [$p(brewpub \mid topic_7)$ = 0.018160818017321258
	brews [$p(brews \mid topic_7)$ = 0.006053606005773753
	crafted [$p(crafted \mid topic_7)$ = 7.722429379421287E-9
	brighton [$p(brighton \mid topic_7)$ = 1.0263658537554439E-14

TABLE 5.2: Example topics from some of the first rounds of topic modelling. Included in the input are all Facebook likes pages and related web pages.

had one, were also included as input documents at this iteration. Statuses, checkins and events were also considered as input sources, but an informal evaluation of the test data from these documents showed that they contained little useful information. Hence they were not included in this iteration. Table 5.2 shows three example topics from an early topic model.

The LinkedGeoData ontology describes what classes are in the data. The thesis use the terms LGD-class and LGD-category interchangeably. Table 5.4 shows some of the LGD-classes. If a query term matched the last part of these URLs the query would return all nodes belonging to this class. In this case match means that the query term and the last part of the URL is exactly the same. The query word “Sport” matched the class “`http://linkedgeodata.org/ontology/Sport`”, but not “`http://linkedgeodata.org/ontology/SportGym`”. One step taken to find these words was to adjust the number of topics the algorithm generates. Too many topics resulted in too topic specific words, while too few topics left out many words. See Table 5.3 for a comparison of the number of topics. The "10 topics" row show that 10 topics was too few topics because each generated topic contained words from several topic-interests. The last row, 30 topics, show that these topics were too special. Both the topics in the table are about specific Facebook-pages; topic 2 is about the pub "Biskopen" and topic 23 is about "Mathallen". However the "15 topics" row demonstrates that the words can be about the same topic and general enough to be used in queries. LinkedGeoData is in English and the topic words were a combination of Norwegian and English words. All the Norwegian words were therefore translated to English using Google Translate API before they could be used in the query.

Running the query selecting only exact matches for words in the topic model gave few POIs. Only some words matched the LGD classes. In order to be able to recommend more POIs I had to tweak the topic model further. "Wines", "Scuba", "Photo", "Brewery" were some of the words from which should have matched the corresponding LGD categories. If you look up "Wines" in the LGD ontology you find "Wine", "Winery", "WineCellar" and "WineShop", but "Wines" don't match these because of the plural form "s". These findings resulted in the adding of a "singular form of the word"-filter. "Scuba" was also a word that should be matching nodes in classes. However the LinkedGeoData had the class "SportScubaDiving" and "ScubaDivingShop" which resulted in the final version of "category-query", which matches "Scuba" with "ScubaDiving" and not "Cuba". In investigating why "Photo" and "Brewery" did not give any matches, I discovered that in the locations I tried there were no registered "Brewery" or "Photo"/"PhotoShop" in that location in LinkedGeoData.

Another tweak to get more POIs was to match topic words with the $rdf : label$ of the nodes. The $rdf : lable$ is a property resource used to provide a human-readable version of a resource's name (W3C, 2004). Many titles of locations contain words which describe what the location is. See Table 6.12 for examples in Section 6.2.2. This resulted in the "label-query" which matches single words in the label/title of the location. Testing these queries in non-English speaking countries showed that the titles were usually in the native tongue of that country. For this reason the query terms had to be translated befor doing the label-queries.

At this stage the Semantic Tourist recommended a lot of points of interest in the selected city. However, an evaluation of these POIs showed that some were locations you would not want to visit. Hence a few tweaks were implemented to remove the unwanted POIs. One of the POIs

10 Topics	food [$p(food\|topic\_1) = 0.029481774395334507$ local [$p(local\|topic\_1) = 0.014082955296835693$ house [$p(house\|topic\_1) = 0.007041350442897237$ medical [$p(medical\|topic\_1) = 0.006616341761691917$ idea [$p(idea\|topic\_1) = 0.003129545596916647$
15 Topics	food [$p(food\|topic\_3) = 0.023482374572870546$ jokke [$p(jokke\|topic\_3) = 0.02315952874228274$ local [$p(local\|topic\_3) = 0.010960598318653899$ joachim [$p(joachim\|topic\_3) = 0.007289427598217145$ job [$p(job\|topic\_3) = 0.0024115133225830337$ dr [$p(dr\|topic\_9) = 0.02267308826173007$ hospital [$p(hospital\|topic\_9) = 0.006140468018761107$ food [$p(food\|topic\_9) = 0.00608430074590709$ episode [$p(episode\|topic\_9) = 0.0045366516561362175$ foods [$p(foods\|topic\_9) = 8.069102527203167E\text{-}4$
30 Topics	food [$p(food\|topic\_5) = 0.06351034501734272$ foods [$p(foods\|topic\_5) = 0.012983577826381616$ healthy [$p(healthy\|topic\_5) = 0.012662239668619015$ health [$p(health\|topic\_5) = 0.010987588737641852$ heartbroken [$p(heartbroken\|topic\_5) = 3.130243532441211E\text{-}20$ få [$p(\text{få}\|topic\_23) = 0.025381156355727786$ mat [$p(mat\|topic\_23) = 0.021763054177707228$ matlagingskunsten [$p(matlagingskunsten\|topic\_23) = 3.3800980327102775E\text{-}12$ mathallens [$p(mathallens\|topic\_23) = 4.4623168992996515E\text{-}13$ mathallen [$p(mathallen\|topic\_23) = 1.6846706187139525E\text{-}15$ photo [$p(photo\|topic\_25) = 0.030416935958438838$ youtube [$p(youtube\|topic\_25) = 0.02656257357121629$ views [$p(views\|topic\_25) = 0.02611935377571408$ mat [$p(mat\|topic\_25) = 0.011324794394081066$ matfestival [$p(matfestival\|topic\_25) = 0.0018451269246064852$

Table 5.3: Example topics from topic models with 10, 15 and 30 topics. "10 topics " created topic with "mixed topic-interests". "30 topics" generated too specific topics, while "15 topics" generated topics which could be used in the queries.

recommended at this stage was hospitals. A hospital is not somewhere you would want to go, especially not on vacation. An investigation of which topic hospital came from showed that it was also related to house, dr, drama etc. (see Table 5.2). which is the TV show House, in which the main characters are doctors working in a hospital. Similar discoveries resulted in the removal of the categories listed in Table 4.3. Another step to remove unwanted POIs was to look at the rank/probability of the topic words giving wrong recommendations. Some of the unwanted topic words were listed among the bottom in the list of topic words. Generating fewer words removed some unwanted points of interest. It was in the investigation of removing

http://linkedgeodata.org/ontology/Sport
http://linkedgeodata.org/ontology/SportScubaDiving
http://linkedgeodata.org/ontology/SportGym
http://linkedgeodata.org/ontology/WaterSports
http://linkedgeodata.org/ontology/Tattoo
http://linkedgeodata.org/ontology/Cosmetics
http://linkedgeodata.org/ontology/Financial
http://linkedgeodata.org/ontology/Campsite
http://linkedgeodata.org/ontology/Restaurant
http://linkedgeodata.org/ontology/Farm
http://linkedgeodata.org/ontology/Skiing
http://linkedgeodata.org/ontology/ArtsCentre
http://linkedgeodata.org/ontology/TourismArtwork

Table 5.4: Examples of LinkedGeoData classes. The Semantic Tourist "category-queries" selects nodes where the query term is equal to (or partly equal to) the last part of the URL

some of the unwanted topic words that the possibility of changing topic modelling library came up. Due to there being multiple controllers in the architecture, it was a small refactoring job to test with another library. After a few quick tests it turned out that Mallet was consistently providing better results, and in less time, than Mahout. Mallet was therefore chosen, at this point, as the topic modelling library to be used.

5.1.2.1 RDF-model

The idea of using a RDF-model to generate event POIs was a result of the elimination of the Facebook-categories related to music interests as input for the topic modelling, and the observation that traditional tourist guidebooks and applications rarely recommend events. It would be a nice supplement to the location-based points of interest.

5.2 Final evaluation

The final evaluation was prepared according to the theories described in Section 5.2.4. In order to address the research questions, a combination of a quantitative measure and a qualitative follow up questionnaire was implemented. The quantitative measure compared the precision of the application against that of a traditional tourist application. The reason for doing quantitative research was because I wanted to test the application for multiple locations with multiple users, and doing this in a qualitative test was considered too time and resource consuming compared to the potential data output. A qualitative follow up was designed to help explaining the quantitative results, especially with respect to the personalization issue.

The main research question inquiries about the feasibility of generating a reliable model of the user's interests using semantic technologies and information from the users social network profile. The evaluation aim for this research question was to gather data on the quality of the personalized recommendations.

The first subquestion inquiries whether Facebook provides sufficient information to give personalized recommendations. The evaluation aim for this question was to gather data on whether Facebook profiles can contain enough information to create a user model that represents the user's interests and preferences. In addition to the testing I also looked at the possibilities of expressing interests and preferences in Facebook, and to what degree user's share this information.

The second subquestion inquires to what extent the general interest of the user can be applied to generate recommendations in a tourist domain. The evaluation's aim was to gather data on the extent to which the application is suited for generating tourist recommendations.

5.2.1 Preparing the testing session

The preparation for the testing session revealed some issues. The application shows *all* the points of interest within a radius of 5 km. This means that if restaurants are among the users POIs, there will be a lot of restaurants to consider for the user. Evaluating a hundred plus restaurants or even just 20 would be tiresome for the user when just having the name, a category and a topic word to go on. Hence I tried to develop a filter for the restaurants, something that would look for any country among the topic words. The hypothesis would be that if a country name would show up in the topic word, then the user could be more interested in this country than others. And he or she would maybe be interested in going to a restaurant with food from this country. However the cuisine property in LGD is not restricted to the origin of the cuisine. There are other "cuisines" like hamburger, pizza and sushi which would have been discarded without consideration by the filter. Consequently this filter was not included as a part of the application.

To adjust the number of POIs the test participant would have to consider, I chose to get 10 random POIs from each category. The idea behind this is that based on the information the application provides and not knowing the exact location of where the user would stay, then you could pick e.g a restaurants at random to be evaluated. This because there are usually more things to consider when choosing where to eat, than the few bits of information the user was now presented with. It would not make a difference to the user at this point, if the bar under consideration were in the East of London or the West because the distance from the hotel is not known. Selecting only 10 POIs was also done with consideration to the test participant. Some of the categories, e.g "restaurant" and "shop", had many POIs, and it would be a very time consuming process for the test participant to evaluate that many POIs in each city. The radius of the search was also decreased to 4 km to decrease the number of POIs further.

The possible answers when evaluating the points of interest were "yes/maybe", "no" and "don't know what this is". "Yes/maybe" meaning "Yes, maybe I would check it out", "No" meaning "im not interested in this point of interest" and "Don't know what this is" is if the test participant could not figure out what sort of location the POI is based on the provided information. The choice of putting yes and maybe under the same option was because of the little information provided about the POI. It also made a strong distinction between the POIs that were *not interesting* at all and those that *could be interesting.*

5.2.1.1 Participant selection

As the application uses the users Facebook data, I asked on Facebook for people who wanted to test my application. Within a day I had 6 people who wanted to participate. Because of the relative short response time, I can assume that these people use Facebook at least once a day. The rest of the testers were respondents to an invitation sent as a Facebook message and as a post in Facebook group I'm a member of. There were 15 test participants in total, 5 pilot testers and 10 final testers. There were 7 female and 8 male between 20 and 35 years old.

The sample validity is defined as whether or not the sampled population reflects the target population of the user group of the application (Bryman, 2008). The selected test participants should be viewed as more of a convenience sample than a probability sample (Neuman, 2011). This is because there is no guarantee that the selected population reflects the target population as this population is unknown. Ideally the number and the variation in the test population should have been higher, but the scope was restricted due to the research projects time constraints.

5.2.1.2 Resource selection

Oslo, London and San Francisco were chosen as the cities to run for the test. Oslo was chosen to demonstrate how the application works with locations in non-English speaking countries. Because it is in Norway many of the test participants were also likely to be familiar with the city. London was chosen because it is a large city with many point of interest, because it is in an English speaking country and because it is a city many Norwegians are familiar with. San Francisco was chosen because it is in a English speaking country and because it was less likely that any of the test participants has any familiarity with it. English speaking countries were chosen to avoid the language barrier for the test participants. Furthermore, the cities were also chosen because of their variation in size.

5.2.2 Evaluation method

In order to evaluate the application I had to find a measure for personalization. A quantitative measure with a qualitative follow up questionnaire was chosen. The quantitative involves letting the user check if he or she would consider visiting the points of interest generated. The test participants evaluated the Semantic Tourist and stay.com, a traditional tourist application. By traditional I mean not personalized, presenting the points of interest in categories as those found in guidebooks. Stay.com has the following categories: hotels, restaurants, attraction, museums, art galleries, shopping, pubs and bars, night clubs, health and beauty, and entertainment. Seven points of interest were randomly chosen from each of the categories at stay.com to be in the final test. If the Semantic Tourist performed better than the traditional application, having a higher precision, it would show that the Semantic Tourist was personalized, because the user was presented with a higher percentage of relevant POIs.

The test participants answers were collected and counted. The responses for each city were kept separate in order to discern a possible difference between the cities. For the same reason the answers from the points of interest generated by the topic model and the RDF-model (eventful.com) were also kept separate. The topic model's points of interest include both label and category queries run against LinkedGeoData as discussed in Section 4.4.1.5.

The responses from the test participant were evaluated with regards to precision, and by a two way analysis of variance of said precision. Precision is one of the simplest measures of the effectiveness of information retrieval systems (Manning et al., 2009). It is the fraction of retrieved documents that are relevant. Precision is defined as:

$$Precision = \frac{true \quad positives}{(true \quad positives + false \quad positives)}$$

Recall is another measure that is often used to measure information retrieval. It is the fraction of relevant documents that are retrieved (Manning et al., 2009).

$$Recall = \frac{true \quad positives}{(true \quad positives + false \quad negatives)}$$

Estimation of recall for a query requires detailed knowledge of all documents in the collection. For this domain it is hard to know what the whole collection, hence it was not chosen as a evaluation measure.

The analysis of variance (ANOVA) is to see if there is any difference in performance between the groups based on one of the variables (Seltman, 2012). A one way ANOVA looks for differences between groups, while a two way ANOVA looks at more complex groupings. It can show the different effect of two separate factors, and the effect of interaction between the factors. This analysis also looks at the statistical significance of the results. That a result is statistically significant means that the results are not likely to be due to chance factors (Neuman, 2011). Using probability theory and specific statistical testing, the statistical significance measure can tell whether the results are likely to be a product of random error in the sample or that the effects are likely to actually appear in the social world. It's important to underline that this only tells about what is likely. It does not prove anything. Statistical significance is usually expressed in terms of significance level: $p = .05$, $p = .01$ and $p = .001$. "The p-value is the probability that any given experiment will produce a value of the chosen statistic equal to the observed value in our actual experiment or something more extreme, when the null hypothesis is true and the model assumptions are correct" (Seltman, 2012). The most common significance levels is $p = .05$ (Seltman, 2012). In this case several two way ANOVA's are used to see if there is a difference between the Semantic Tourist and stay.com and if the results varies according to location, using a significance level of $p = .05$.

In addition to this quantitative measure of personalization, a follow up questionnaire was given to the user. The questionnaire was designed to cover the parts about personalization that the

Follow up questions
1. Do you recognize yourself in the words in the topic list?
2. Is there some topic that surprise you? In what way?
3. Is there anything you have ”liked” on Facebook that are not represented in this topic?
4. Do you find an app like this is useful?
5. In what situations can you imagine that this app is useful?
6. Have you visited Oslo before? If yes, how many times?
7. Have you visited London before? If yes, how many times?
8. Have you visited San Francisco before? If yes, how many times?
9. What types of things do you look for when deciding on one week's vacation?
10. How often do you use Facebook?
11. What types of sites do you like on Facebook?
12. Do you have any other comments?
13. Were there some categories in my application that you did not recognize yourself in?
14. If you answered yes to previous question: What do you think is the reason you do not recognize yourself in the categories?

Table 5.5: Follow up questions.

quantitative test could not cover and collect qualitative additional information about the user's perceptions about the system. It included the questions presented in Table 5.5. Questions 1 and 13 could give an indication wether there was something wrong with the topic model. The 3rd and 11th question give an indication on wether the information collected from Facebook is the right to use. Question 9 gives an indication about what the user is looking for while on vacation. A qualitative analysis comparing the test participants individual results and the follow up questions was conducted for further investigation the application.

5.2.3 Pilot test

The pilot test was conducted with five test participants. Among the tree first test participants, two indicated that they did not share much personal info on Facebook. Hence two more test participants were added to the pilot test. Evaluation of the pilot test revealed that more categories should be included in the "excluded categories"-list for input documents to the topic modelling algorithm. In the investigation following the pilot test the subjects talked about their Facebook usage. The conversation with the test participant indicated that statuses updates and checkins could be added as input documents to the topic model. The test participants indicated that statuses updates could be included because they used statuses updates to share thing of interest, and update about things that were important to the them. They also indicated that checkins for them would be related to a positive experience in most cases. Hence a decision to add statuses updates and checkins was made on the basis of the conversation with the test participants. The decision was made despite the initial assessment to leave these out.

The pilot test, like the testing during development, ran with the same number of output topics from the topic modelling for everyone. It did not consider the number of input documents. The results from the pilot test showed that this number seemed to be too large for the users with a smaller amount of input documents. An adjustment was implemented to deal with this problem. The idea was that the number of output topics should be proportional to with the number of input documents. Different variations were tried, but it seemed that dividing the number of input documents by 5 and putting an upper limit of 40 could be right. Additionally the number of topic words from each topic was decreased from 5 to 3 for the same reason explained above. In the final test these was the parameters for picking out the number of topics to generate.

5.2.4 Test session

The test sessions began with the user receiving a letter of consent (see Appendix B) to make sure the user knew the circumstances of the test. A brief explanation of the procedure and the intention of the test followed to make the user aware of this. The user was then asked to try the application with the test cities. Directions were given to help the user understand the category system. The extra view of the topic model was also shown to the user to help explain the idea behind topic modelling, and so the user could answer the follow up questions. When the user was finished exploring the application, he or she was given the documents with the points of interest. The document contained the points of interest of the Semantic Tourist and stay.com grouped by city and POI type (location and event). An explanation was provided to help the user evaluate the POIs as intended (see Section 5.2.1).

5.2.4.1 Hardware and software

The evaluations during the development and final test was run on the same server machine. The Android device was the same for most of the test participants. However some had their own Android device and the application was installed and tested on their device.

Server machine: Multicom Xishan W150E, memory: 8 GB, cpu: Intel® Core^{TM} i7-3720QM, disc: 128 GB SSD.

Client device: A Samsung Galaxy S3 running Android version 4.1.2, and in some cases the user's own Android device.

5.3 System evaluation

The aim of the project was to create a prototype of the system that would demonstrate the techniques and technology used. The scope did not include good user interface, session handling, security or other features necessary to put the application in the Google Play store. However, this section presents some informal assessment of the system related to the "Design evaluation" guideline presented in Section 3.1.1.

The usability of the application was something that the test participants made comments about during the test session. Comments about the user interface indicated that more work could be done in the design of the screens. One person mentioned that she liked the simplicity of the application. She liked that you open the app, enter the destination, wait for the result and you get all the different recommendations in one app.

The performance time of loading the points of interest was another feature that got some remarks. The topic modelling algorithm uses minutes to generate the topic model. Also there are potentially over 200 queries run against the LinkedGeoData SPARQL endpoint which take time. Consequently running the application for a new user could take several minutes. The topic model is however only updated once a day, saving some minutes after the first time you run it. The runtime is not something that can easily be improved because of the time it takes to run the topic model algorithm. But some of the test participant commented that this would be time you normally would spend on browsing. Consequently the long loading time was acceptable as long as the application provided good recommendations.

The application is at a stage of early high fidelity prototype. As mentioned, there are features that must be implemented before the application can be deployed. Additionally there are features that can improve the results of the application (see Section 7.2)

Chapter 6

Results and discussion

6.1 Results

This section will present the results of the final evaluation of the Semantic Tourist application described in Section 5.2. The results are analysed with the methods described in Section 5.2.2. First the results of the qualitative analysis will be presented. These are findings from the follow up questions and the individual test participant results. Secondly we will look at the results of the quantitative analysis comparing the two tools and the two models within the Semantic Tourist application.

The research questions that we started with was this: Is it possible to generate a reliable model of a user's personal interests using his or hers social network presence and semantic technologies? We also had two sub questions which we will look at: Does Facebook provide sufficient information to do this? and to what degree can the resulting model be used to generate personalized recommendations in the tourist domain?

6.1.1 Qualitative analysis

Figure 6.1 shows a comparison between the Semantic Tourist and stay.com for each test participant. The bars in the chart represent the individual users' mean precision for the three cities. It shows that three test participants had (relatively) higher precisions than the rest, above 70%. The test participant with the highest mean precision also answered *"Very. I think it is very much relevant to my interests, especially music interests. But also other things like*

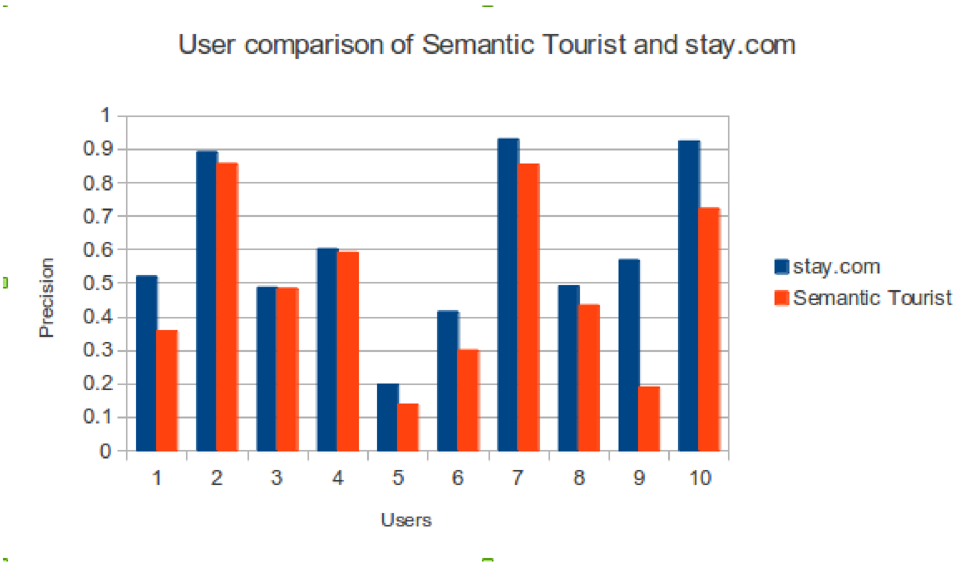


Figure 6.1: The diagram shows a comparison between the Semantic Tourist and stay.com for each test participant

sport and museums" to the second follow up question. Table 6.1 shows the topic model of the test participant, and Table 6.2 shows a section of the points of interest that were recommended to the same participant in San Francisco.

tegl, pub, aug	voss, myrkdalen, ligg
hug, cycling, relationship	bredbånd, itavisen.no, internett
game, players, player	world, cup, boats
kilroy, tilby, verden	prezi, edge, animate
piano, website, strings	people, social, facebook
spotify, service, customer	aftenposten, største, norges
psykologi, psykologisk, artikler	naxos, music, library
helse, kunnskap, temaer	kvarter, arrangementer, pir
business, local, tilbud	free, html, download
karate, japanese, strikes	universitetet, bergen, studenter
maria, norge, vilvite	høyres, studenterforbund, samfunn

Table 6.1: Topic model of test participant with above 70% precision.

Two test participants had less than 20% precision with the Semantic Tourist. One of these also had below 20% precision in stay.com. This person had remarks about topics he did not recognize and that he felt it was very hard to decide if he was interested in the POIs based on the little information provided in the test. The other person below 20% precision commented that he recognized the topic model "moderately" and had remarks about missing topics. However he also commented that he rarely "likes" things on Facebook. Table 6.3 show the topic model of one of the participants with low precision, and Table 6.4 show a section of the same participants recommended points of interest in San Francisco.

y/n/?	Title/label	LGD category	Query word	Query
y	Clooney's Pub	Bar	Pub	l
n	Bing-Kong Tonc Free Masons	PlaceOfWorship	Free	l
y	Golden Gate Community Church of the Nazarene	PlaceOfWorship	Community	l
y	Chinese Community Church	PlaceOfWorship	Community	l
y	Community Baptist Church	PlaceOfWorship	Community	l
y	Seventh Day Adventist Japanese Church	PlaceOfWorship	Japanese	l
y	San Francisco Evangelical Free Church	PlaceOfWorship	Free	l
y	Mission Branch Library	Library	Library	c
y	Far West Library for Educational Research and Development	Library	Library	c
y	Mechanic's Institute Library and Chess Room	Library	Library	c
y	Western Addition Branch Library	Library	Library	c
y	Erik Erikson Library	Library	Library	c
y	Golden Gate Valley Branch Library	Library	Library	c
y	Presidio Branch Library	Library	Library	c
y	William E Colby Memorial Library	Library	Library	c

TABLE 6.2: San Francisco points of interest from test participant with above 70% precision. 15 POIs of the full Table D.1

The example points of interests displayed in Table 6.2 and Table 6.4 show the participants was recommended different POIs. There was however one exception that both was recommended places of worship retrieved by the "Label-query". In Table 6.2 the query words used to select the POIs were "Free, Community and Japanese" while in Table 6.4 "Free, Chi and Day" were the query words. One participant considered it a place he would visit, while the other did not.

Figure 6.1 also shows a high correlation (r=0.8) between the results of the two applications. Meaning that a user with a high precision in one application probably has a high precision

page, official, facebook	team, switzerland, cup
outdoor, products, mammut	sheldon, leonard, penny's
devold, krav, plagg	kvarter, huset, akademiske
world, garage, skis	och, bergen, för
norge, intersport, dag	dry, les, chamonix
google, chrome, chi	station, radio, broadcasts
åsane, storsenter, gratis	völkl, ski, freeski
world, freeride, north	power, days, movies
varen, landets, priser	åsane, skyte, shooting
software, gnu, free	bergen, student, studenter
skisenter, voss, hemsedal	jakt, fiske, norges
und, gore-tex, die	film, tornatore, cinema
hjortesenter, norsk, hjort	shotguns, rifles, howa

TABLE 6.3: Topic model of test participant with below 20% precision.

y/n/?	Title/label	LGD category	Query word	Query
n	Metreon	Cinema	Cinema	c
n	Sundance Kabuki	Cinema	Cinema	c
n	Bing-Kong Tonc Free Masons	PlaceOfWorship	Free	l
n	Chi Sin Buddhist and Taoist Association	PlaceOfWorship	Chi	l
n	Seventh Day Adventist Japanese Church	PlaceOfWorship	Day	l
n	San Francisco Evangelical Free Church	PlaceOfWorship	Free	l
n	Central Seventh Day Adventist Church	PlaceOfWorship	Day	l
n	Philadelphian Seventh Day Adventist Church	PlaceOfWorship	Day	l
n	Seventh Day Adventist Tabernacle	PlaceOfWorship	Day	l
n	Admission Day Monument	Monument	Day	l
n	City College of San Francisco Chinatown North Beach Campus	School	North	l
n	Horace Mann Academic Middle School	School	Academic	l
n	Saint Marys Chinese Day School	School	Day	l
n	Hillwood Academic Day School	School	Academic	l
y	Radio Shack	Electronics	Radio	l

Table 6.4: San Francisco points of interest from test participant with below 20% precision. 15 POIs of the full Table C.1

in the other and vice versa. Only one participant had a big difference in precision, about 35 percentage points in favour of stay.com

In the follow up questions (questions 1-3) all the participants recognized the topics from the topic model, and 4 of 10 answered that they recognised it without any remarks. Three mentioned that they missed categories that were intentionally excluded from the topic model (see Section 4.4.1.2), so for the purpose of this testing, 7 out of 10 approved the topic model as a representation of their interests expressed on Facebook. The three that did not recognize all the topics were also the three test participants with the lowest precision in the Semantic Tourist application. Among them, one mentioned a specific topic (beer brewing) that he missed in the POIs. Additionally he did not know where the topic words “irish”, “bank” and “metro” came from. The second person who did not recognize all the topics commented that he had used Facebook for a long time and that he had changed over the years. The last person who made remarks about the topics, commented that there were several topics he did not recognize, e.g. “och-bergen-for” and “dry-les-chamonix”.

On the related question; “are there some of the categories you do not recognize?”, only 8 replied. However 4 of the 8 answered that there was nothing they did not recognize; two mentioning specific topics that they did not recognize, one saying that it was most likely because he had been using Facebook for many years, and one could not remember which topics he did not recognize. The latter was also the person with the highest precision for Semantic Tourist. The participants with second and third highest score for the Semantic Tourist were among those who answered that there was nothing they did not recognize.

The answers to the $11th$ follow up question show that all of the participants share pages about interests and hobbies. 9 of 10 share favourite music/TV shows/books/movies etc. Five share pages for a "good cause" and two commented that they share little personal information on Facebook.

When asked about situations in which the application could be useful, the participants answered e.g; "in planning vacations in unknown cities", when "you do not know what to do in an unknown city and do not want top 10 attractions from the guide book", "in situations where you do not want to spend a lot of time researching what to do", "and to find events in a city you do not know or in your home town". Additionally 6 gave positive remarks about the events POIs. One user was especially excited about the application. He commented that he was planing a trip to Thailand. After trying the application however, he considered changing the destination to San Francisco because of the recommendations the application gave.

6.1.2 Quantitative analysis

This section presents the quantitative results of the evaluation. In the following, when I talk about "tools" it will refer to Semantic Tourist and stay.com. When locations are discussed, they are the test cities London, Oslo and San Francisco.

SemT	London	Oslo	San Francisco
mean	.534	.533	.411
std. deviation	.276	.312	.245

Table 6.5: Mean precisions and standard deviations for the Semantic Tourist

stay.com	London	Oslo	San Francisco
mean	.589	.615	.605
std. deviation	.257	.226	.248

Table 6.6: Mean precisions and standard deviations for stay.com

The results from stay.com have little variation in precision between the cities, all with around 60% precision. The Semantic Tourist has a lower mean of about 53% in London and Oslo, and 41% in San Francisco. The statistics are shown in Table 6.6 and Table 6.5

The graph in Figure 6.2 shows the differences in precision between Semantic Tourist and stay.com for the different locations. This analysis includes both POIs generated based on the topic model

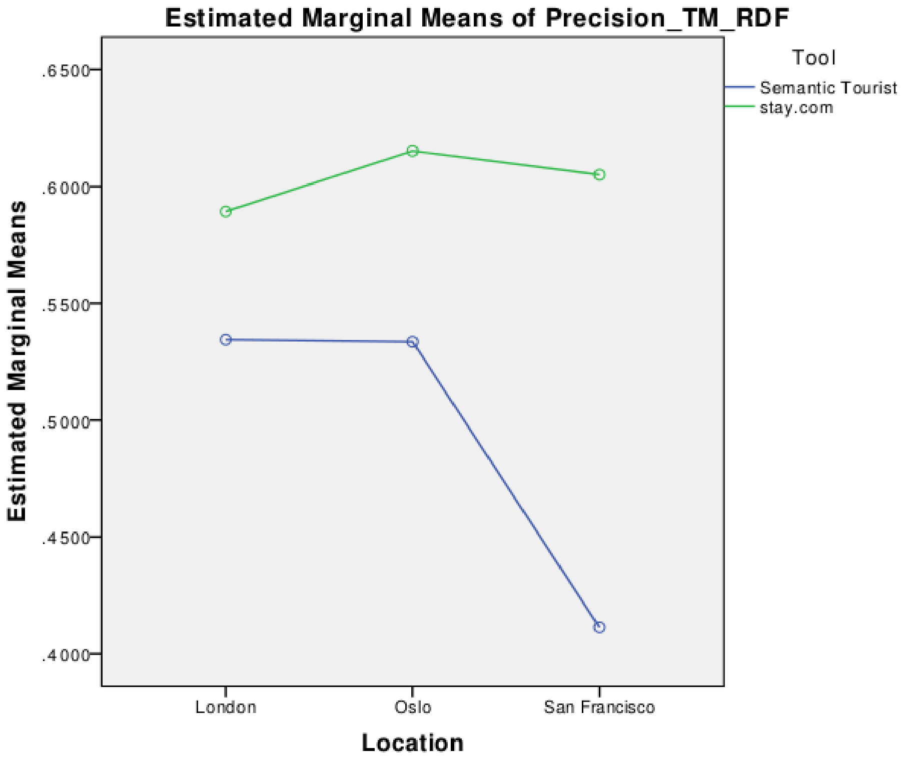
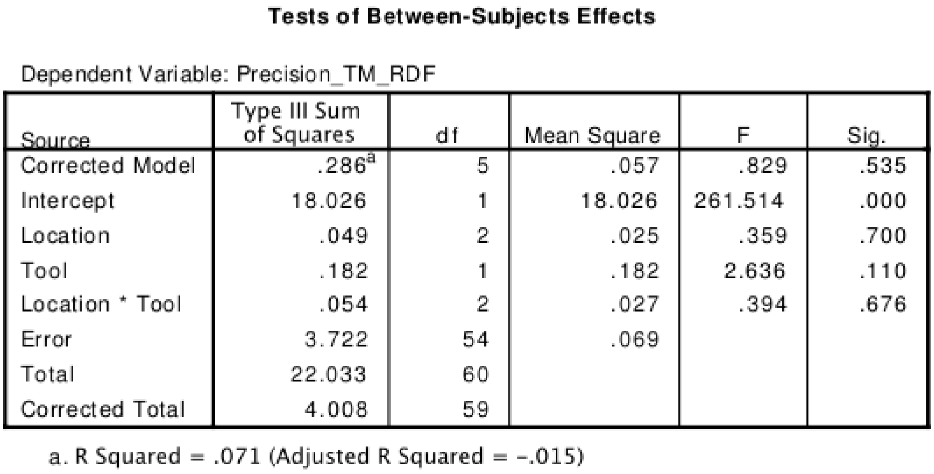


FIGURE 6.2: The mean precision of points of interest in Semantic Tourist (topic model and RDF-model) and stay.com

and the RDF-model. In London, the Semantic Tourist has 53% precision about 5 percentage points lower precision than stay.com. In Oslo it has about 53% precision is 8 percentage points lower, while in San Francisco the precision of Semantic Tourist dropped to around 41%, about 19 percentage points lower than stay.com.



Tests of Between-Subjects Effects

Dependent Variable: Precision_TM_RDF

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.286[a]	5	.057	.829	.535
Intercept	18.026	1	18.026	261.514	.000
Location	.049	2	.025	.359	.700
Tool	.182	1	.182	2.636	.110
Location * Tool	.054	2	.027	.394	.676
Error	3.722	54	.069		
Total	22.033	60			
Corrected Total	4.008	59			

a. R Squared = .071 (Adjusted R Squared = -.015)

TABLE 6.7: Statistical summary of precision performance of Semantic Tourist (topic model and RDF-model) and stay.com

However, when we analyse the numbers as shown in Table 6.7, we see that the differences are not statistically significant for either location or tool, nor the interaction between the two when using $p = .050$ as significance level as described in Section 5.2.2. Even though Semantic Tourist shows a lower precision in the test, we must conclude that there is no statistically significant difference between the precisions of Semantic Tourist and stay.com.

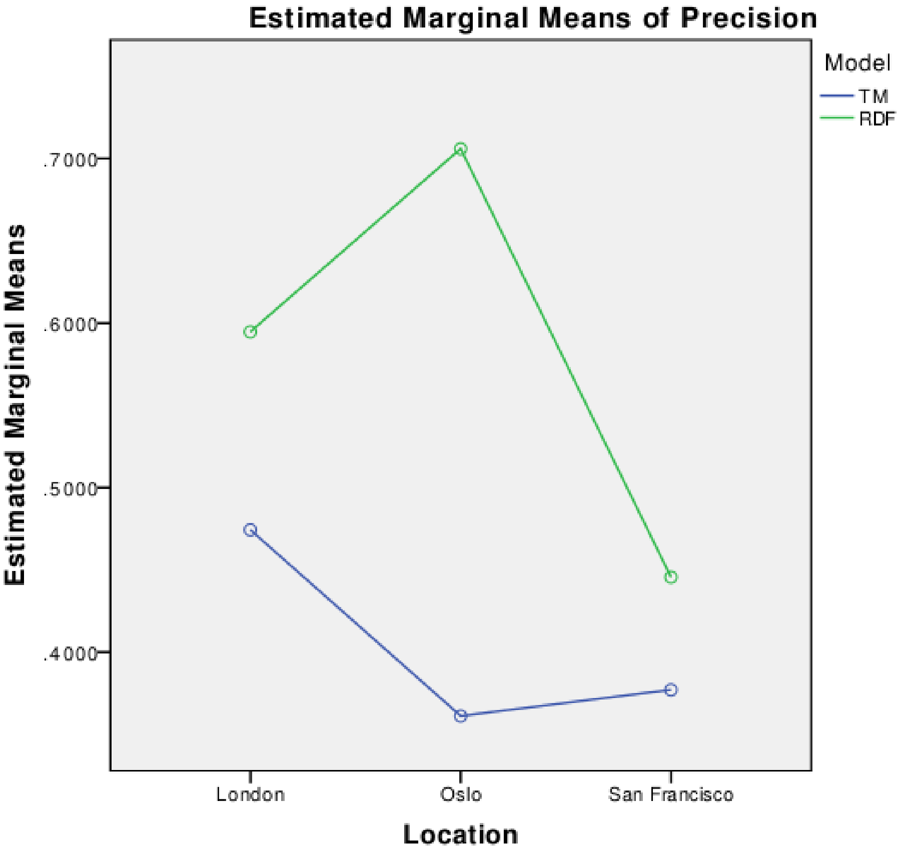
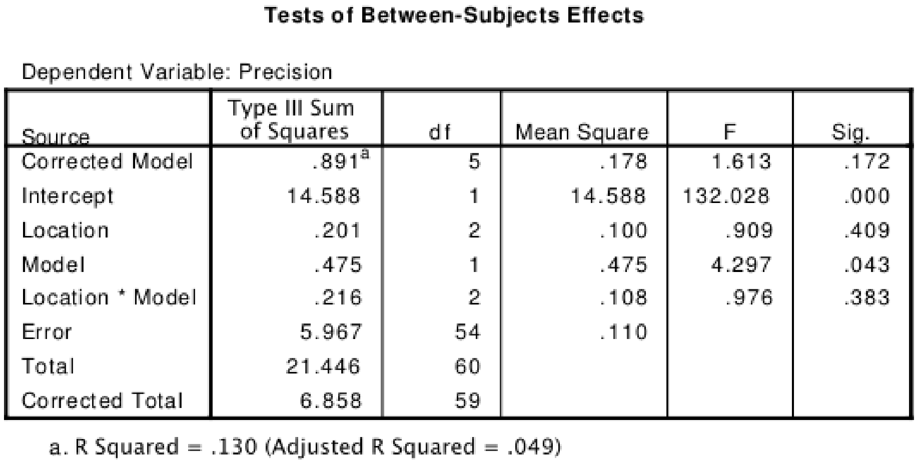


FIGURE 6.3: The mean precision of points of interest from Semantic Tourist topic model (LinkedGeoData) and RDF-model (eventful.com)

There are two different models used in the Semantic Tourist, built using different techniques. Figure 6.3 shows the comparison of the precision means for the topic model and the RDF-model. It shows that the precision of the event POIs are higher in all cities. The results, as shown in Table 6.8 confirms that there is a statistical significance ($p = .043$) for model, but not for location or the interaction between location and model. We can therefore conclude that the model is a statistically significant factor in describing the difference between the results.

Figure 6.4 shows the analysis of the results from the topic model points of interest (POI) compared with stay.com. The topic model POIs contain results from the label and the category query discussed in Section 4.4.1.5. From the graph in Figure 6.4 we see that the POIs generated



Tests of Between-Subjects Effects

Dependent Variable: Precision

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.891[a]	5	.178	1.613	.172
Intercept	14.588	1	14.588	132.028	.000
Location	.201	2	.100	.909	.409
Model	.475	1	.475	4.297	.043
Location * Model	.216	2	.108	.976	.383
Error	5.967	54	.110		
Total	21.446	60			
Corrected Total	6.858	59			

a. R Squared = .130 (Adjusted R Squared = .049)

Table 6.8: Statistical summary of precision performance of Semantic Tourist comparing the topic model and RDF-model results.

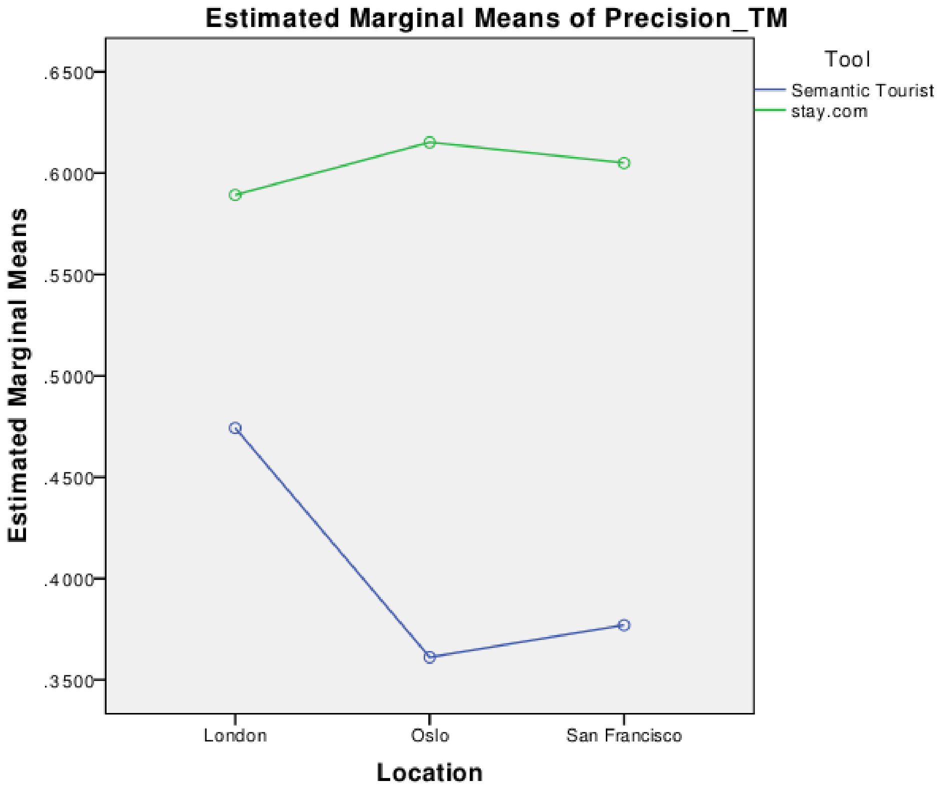
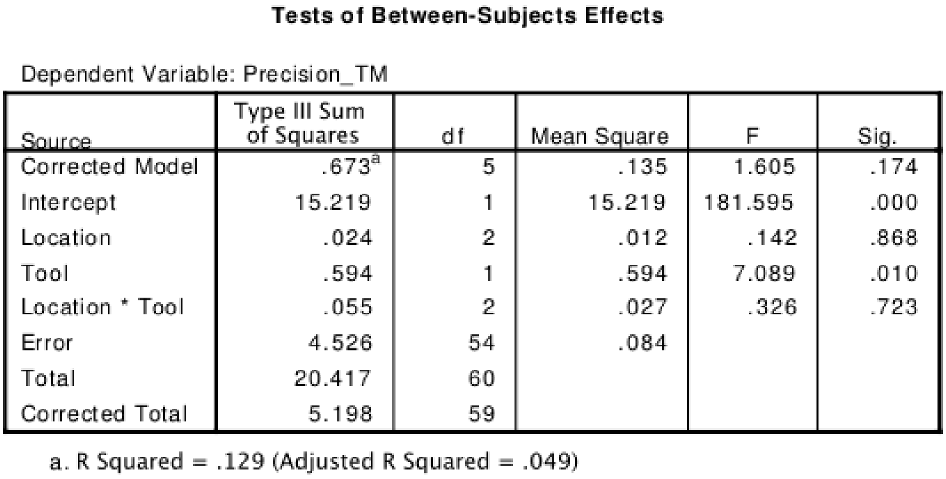


Figure 6.4: The mean precision of points of interest recommended by the topic model in Semantic Tourist and stay.com

by the topic model have lower precision than stay.com. This model has best precision in London ($\approx$ 47%) and worst in Oslo ($\approx$ 36%). Compared with stay.com there is a difference of 9 percentage points in London, 25 in Oslo and 23 in San Francisco.

Table 6.9 shows us that the differences are not significant for either location (p = .868) or



Tests of Between-Subjects Effects

Dependent Variable: Precision_TM

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.673[a]	5	.135	1.605	.174
Intercept	15.219	1	15.219	181.595	.000
Location	.024	2	.012	.142	.868
Tool	.594	1	.594	7.089	.010
Location * Tool	.055	2	.027	.326	.723
Error	4.526	54	.084		
Total	20.417	60			
Corrected Total	5.198	59			

a. R Squared = .129 (Adjusted R Squared = .049)

Table 6.9: Statistical summary of precision performance of the topic model in the Semantic Tourist and stay.com

the interaction between location and tool ($p = .723$). There is however a statistical significant difference between the tools ($p = .010$).

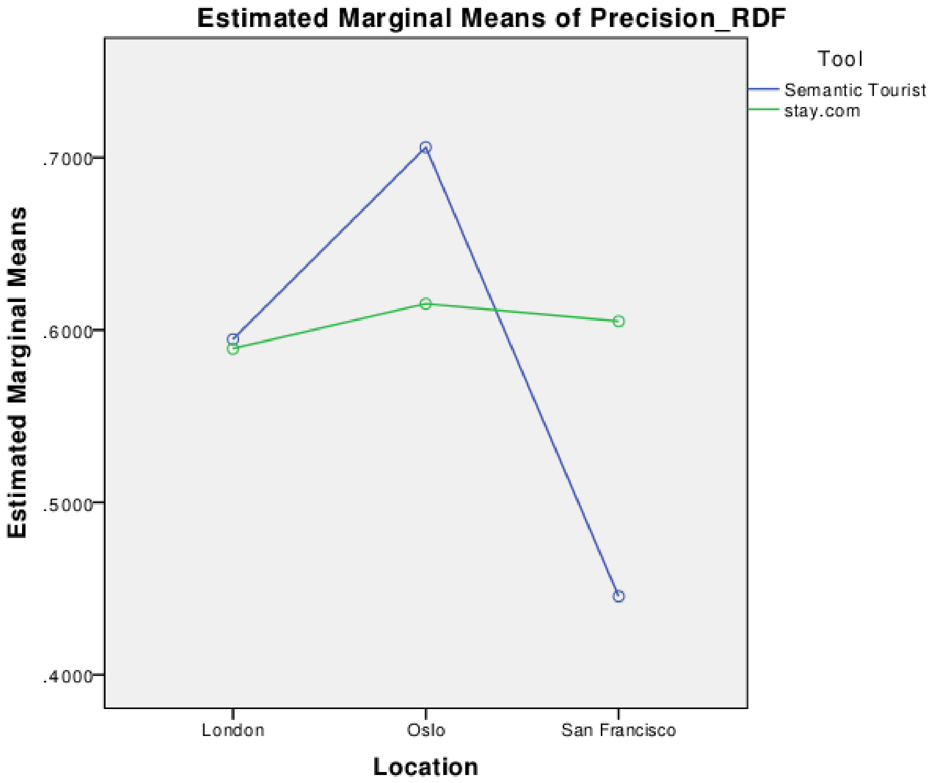
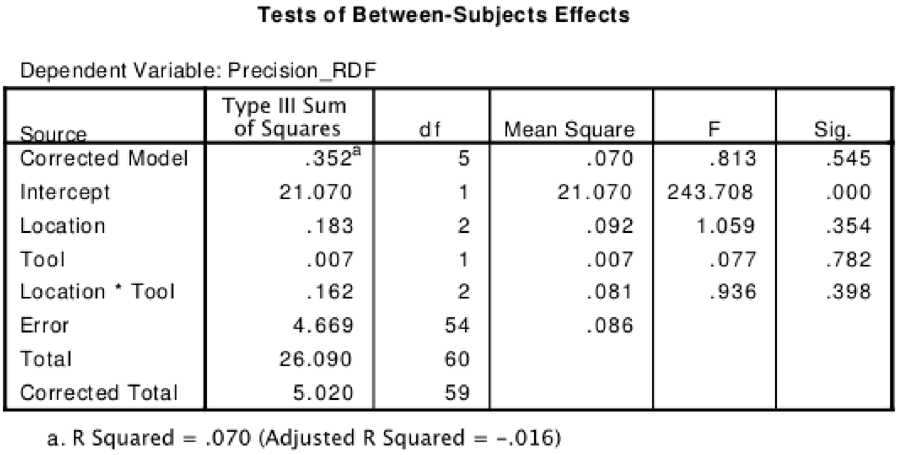


Figure 6.5: The mean precision of points of interest in Semantic Tourist (RDF-model) and stay.com

The RDF-model had better precision than the topic model. Therefore an analysis of the RDF-model (events POIs) compared to stay.com is warranted. Figure 6.5 shows that the Semantic

Tourist RDF-model (59%) has a one percentage point higher precision than stay.com in London. In Oslo the precision is ($\approx$ 70%) which is 9 percentage points higher than stay.com. While in San Francisco stay.com has a 16 percentage points higher precision than the Semantic Tourist (44%). However, the results, as shown in Table 6.10 show us that there is no significant difference between the tools (Semantic Tourist RDF-model and stay.com) ($p = .782$), locations ($p = .354$) or interaction between tools and location ($p = .398$).



Tests of Between-Subjects Effects

Dependent Variable: Precision_RDF

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.352[a]	5	.070	.813	.545
Intercept	21.070	1	21.070	243.708	.000
Location	.183	2	.092	1.059	.354
Tool	.007	1	.007	.077	.782
Location * Tool	.162	2	.081	.936	.398
Error	4.669	54	.086		
Total	26.090	60			
Corrected Total	5.020	59			

a. R Squared = .070 (Adjusted R Squared = -.016)

Table 6.10: Statistical summary of precision performance of Semantic Tourist (RDF-model) and stay.com

To summarize: Although stay.com has an overall higher precision than the Semantic Tourist, this difference is however not statistically significant. Separating the two models behind the Semantic Tourist however shows that there is a statistical significant difference between the two models, where the RDF-model has the higher precision than the topic model. There is also a statistical significant difference between the topic model and stay.com in stay.com's favour. The comparison of the RDF-model against stay.com show that the RDF-model has higher precision in London and Oslo, but lower in San Francisco. These results however are not significant.

6.2 Discussion

The qualitative analysis show that the participants' individual results vary. Three participants had a precision above 70% while two had a precision below 20%. The precision measure for the overall performance of the two tools show that the personalized Semantic Tourist performed a little worse than the traditional tourist application stay.com. However the two way analysis of variance show that the results are not statistically significant. Separately comparing the topic model and the RDF-model against stay.com show that only the topic model performed significantly worse than stay.com, while the difference was not significant between RDF-model and stay.com. This section will discuss the results and point to features that might explain them.

The project was based on the assumption that people would want to do other things than visit traditional tourist attractions while travelling. Based on this assumption, looking at the users general interest seemed like a good starting point for recommending "untraditional"/personalized points of interest. Making personalized categories based on the user's interest would be a good thing because different people want to see and do different things. However, the results show that this assumption is not entirely correct. Some want to see the typical tourist attractions while others do not. The results also show that there is no significant difference between stay.com and the Semantic Tourist. This indicates that they are as likely to visit the traditional attractions as they are visiting the the personalized POIs recommended by the Semantic Tourist.

6.2.1 Facebook as a source

The results show a varying degree of success for the Semantic Tourist. However 7 of the 10 participants answered that they recognized themselves in the topic model based on their Facebook profile. While only two participants had low precision and remarks about there being at least one topic that they did not recognize. 9 of the 10 participants answered that they share information about interests/hobbies and favourite music/books/movies etc. The results from the POIs recommended based on the RDF-model and the participants positive remarks about the recommended events also indicate that Facebook can provide sufficient information about the user's interests. This indicates that the semantic technologies used in this study combined with Facebook as a source of user information, is able to generate a reliable model of the user's interests, which they recognize. Despite the fact that there are needs for improvement with the approach, the findings of the study correspond to the findings of Shapira et al. (2012), which indicate that Facebook can be a suitable source of for information about the user's interests.

Looking at the information that the test participants had in their profile, there are some features that are worth commenting on. Likes listed under Music, Interests and Activities can give a good basis for recommending events. Events may be more attractive than the other POIs, because a persons likelihood of going to an event is less dependent on the location. The fact that all the participants answered that they share pages about interests and hobbies, and 9 of

Event: UK Pink Floyd Experience – key term: Pink Floyd
Event: Patti Smith and Her Band – key term: Patti Smith
Event: The FA Cup Final – key term: Soccer
Event: SPIN London – key term: Bicycling
Event: Adult Karate Class at One Martial Arts – key term: Karate
Event: Neo4j Tutorial – London – key term: Java

Event: CPD Event: How Research Informs and Improves the Process and Outcomes of Restorative Justice – key term: Justice
Event: Becoming Your True Self: Resolving Traumatic Entanglements, LONDON. – key term: Faith No More

Table 6.11: Examples of event points of interest. The key term is selected from the RDF-model. The first POIs are examples of good recommendations, while the last two are examples of "bad" recommendations.

10 answered that they share pages about favourite music, combined with the users' positive comments about the events, strongly supports this assumption. The users' commented that they found the events useful in an unknown city as well as in their home city. The results from the quantitative testing of the RDF-model also supports this. The precision of these results can be influenced by the fact that these recommendations were based on interests from Facebook listings from the similar domains. Music and activities related likes were used as base for recommending related events. Table 6.11 show examples of "good" recommendations (the top events) and "bad" recommendations (bottom two events). The ontology suggested in Section 4.3.4 for the RDF-model, could be used to prevent using band names for recommending events other than concerts.

As showed in Section 4.3.1 Facebook profiles can contain a lot of information. However, users provide varying amounts of information in their Facebook profiles. The answers from the follow up questions show that a few of the test participants share little personal information on Facebook. When this is the case, there is not enough information, and it is not possible to generate personalized recommendations. The results show that one of the two with lowest precision also commented that he share little personal information. This mildly indicates that the initial assumption that users have to share a certain amount of information in the social network for the application to be able to give personal recommendations, was right. For users like this, explicit information user model generation could be better.

The answers from the follow up show that the test users recognise their Facebook profiles in the topic model. However the precision for the POIs generated by the topic model is below 50%.

The following can explain how human mistakes in creating Facebook pages and filtering gaps contribute to the low precision.

The pre-processing filter for the topic modelling tries to remove information that is not in any way descriptive of the page (see Section 4.4.1.1). There is no guarantee however that the people who created these pages put the right information under the right attributes. One example from the test showed that a business had used the description field for listing phone numbers and opening hours for their shops, despite that Facebook provides separate fields for this information. Another example from the test showed that the same concept can be listed several times under different categories. Consequently, if a TV show has been listed under a different category, e.g "Interest", the filter will think it is an interest and include it in the document collection. This would then result in a topic, e.g topic 2 in Table 5.2 being included despite the filter that is supposed to exclude TV shows from the input. The application would hence wrongfully recommend hospitals. As of May 2013 it looks like Facebook has put more restrictions on the categorization, thus this might be less of a problem now.

5 of 10 answered that they share pages about a "good cause". A "good cause" was not specifically defined in the questionnaire. However the Facebook categories "Community organizations", "non-governmental organizations", "charity organizations", "non-profit organizations" etc. could fall into the "good cause" category. Despite the fact that two of the users with high precision answered that they liked "good causes", these pages do account for some of the POIs that the users did not want to visit. Although no clear indication can be drawn from the results as to how the "good cause" pages work as a basis for recommendations in the tourist domain, one could guess that just because someone supports the fight against cancer it does not mean that they wan to visit a cancer research center or a hospital while on vacation. It could mean though, that they would be more likely to attend an event that is specifically aimed at people supporting this cause, such as a rally to support cancer research. Perhaps these pages should be included in the RDF model instead, just like music interests just be used to search for related events.

6.2.2 Topic model

The results from the quantitative evaluation showed that the topic model had a mean precision between 35 and 50%. The ANOVA analysis showed that this part of the Semantic Tourist performed significantly worse than stay.com. It also performed significantly worse than the

RDF-model in the Semantic Tourist. Although 3 of the 10 test participants had above 70% precision with the Semantic Tourist, the overall results show that there is a need for improvement in how the topic model is translated into specific recommendations. The higher precision of the RDF model, and the fact that all the testers recognized their Facebook profile in the topic model are both clear indications that such an improvement is achievable.

There are several possible explanations of the mediocre results of the topic model. A simple explanation could be that the model just isn't an accurate model of the user's interests, but the fact that all the users recognized the topic model, and that 7 out of 10 recognized all the topics in the model indicates that this is not the case. Additionally 4 of the 8 responding, commented that there were no topics they did not recognise. These findings suggest that topic modelling can be used to generate a model that the users recognize. Even if the findings are not strong they point in the same direction as earlier research in that topic modelling of the user's social network presence can be used to model the user's interests.

Not enough information in the Facebook profile could be another explanation. This problem is also discussed in Section 5.2.3 along with the changes that were made to solve this problem. There were however no subjects in the final evaluation that had below 60 input documents to the topic model. The results from the participant with low precision and comments that he had not provided much personal information, and the results from the pilot test, indicate that a certain amount of personal information is needed.

The "label-queries" in some cases generate POIs that are related to the user's interests. The POIs from Table 6.12 are examples from the test where the users answered that they wanted to visit the location. However, in many cases these "label-queries" generated POIs that the users were not interested in. This was a known problem with the "label-queries". The examples from Table 6.2 and Table 6.4 demonstrate one part of the problem. An explanation for the different responses could be in the query terms used in the selection of the POIs. The word community could be a little more related to a place of worship than the more general word day. This could be the reason why this person would visit the place of worship. Another explanation of the different responses could be that one was interested in these type of locations while the other was not.

The problem with the "label-query" can be explained by the way the topic word is used directly in the query. The query does not consider the semantics of the word and returns POIs where the query term matches one word in the title/label. Positive examples from the tests are the

POI: “John Foleys Dueling Piano Bar” - Category: Nightclub (topic: “Piano”) (l)
POI: “Department of Coffee and Social Affairs” - Category: Cafe (topic: “Social”) (l)
POI: “The Japanese Canteen” - Category: Cafe (topic: “Japanese”) (l)
POI: “See Jane Run Sports” - Category: Shop (topic: “Run”) (l)
POI: ”The Heart of Hatton Garden” - Category: Jewelry (topic: ”Garden”) (l)
POI: ”Angel Food & News” - Category: Newsagent (topic: ”Food”) (l)

Table 6.12: The first four POIs are examples of good label matching, while the two last are examples of unrelated label matches. The word listed as “topic” was the query term.

POI: “Holland and Barrett” - Category: HealthFood (topic: “Food”) (c)
POI: “Bagel Street” - Category: FastFood (topic: “Food”) (c)

Table 6.13: Examples of results from the query term “food”. This is an example of a category that is to general. The word listed as “topic” was the query term.

topic words “wines” which matched the shop title “food and wines” and “Piano” which matched “John Foley’s Dueling Piano Bar”. See other examples in Table 6.12 There are however also negative examples: the topic word “wines” also matched a parking lot called the “Great wines parking” and “design” matched “Chelsea College of Art & Design”. See other examples in Table 6.12.

The category queries also generated POIs which the user did not want to visit. In contrast to the label queries these POIs are in some cases not specific enough. The topic model generates a topic containing the words: healthy, food and local. However the query only used the word “food” and was not able to tell the difference between the sub categories of food for example fast food and health food. See example in Table 6.13.

The topic model extracts the right interests; applying them as query terms against LinkedGeoData directly may however not have been the best solution. Some of the semantics are lost between the topic words and using each word as a query term. A possible solution to this would be to create an ontology on top of the topic model to improve the transition of the semantic meaning between the topic model and data source (LinkedGeoData). LinkedGeoData does not state that there is any difference between the two food categories from Table 6.13. An ontology could contain statements that make it possible differentiate between the e.g the two food categories. A feature similar to Michelson and Macskassy (2010) that use Wikipedia’s category system to leverage term ambiguity can also help application of the topic model to the tourist domain in the Semantic Tourist. The application could use the other words the topic to determine the right semantic meaning of the current word. Another feature that can improve the precision of the topic model is more preprocessing of the documents going into the topic model.

POI: "Chesterton" - Category: EstateAgent (topic: "Est") (c)
POI: "Lloyds" - Category: Bank (topic: "Ban") (c)

Table 6.14: Examples of results of query terms that should not have been used. The word listed as "topic" was the query term used to recommend this POI.

Lee et al. (2011) performs a tf-idf weighting removing words falling below 0.5% an retaining documents containing at least one top 0.5% of words. It is not certain that this method is desirable in the Semantic Tourist because you risk loosing words that are related to interests that the user is not that active in promoting in his or her profile. However if it could help in removing more of the unwanted topic words, weighting could be a useful feature.

A few of the negative results are caused by the fact that the filtering of the topic words was not good enough. It missed words like "Ban" or "Est". These words could come from a semantically incorrect translation combined with adding singular forms of nouns. These particular words match "Bank" and "EstateAgent" (see Table 6.14).

The results show that the difference between the applications are not significant. The results from stay.com show that most people want the "typical" tourist recommendations, and the results from the Semantic Tourist show that it can create personalized categories and find points of interest based on the user's Facebook profile. This could indicate that a combination of the two approaches would lead to better results. The high correlation between the tools for each of the users also support this assumption.

6.2.3 Tourist domain/location

Another aspect to the performance of the topic model is that some of the topic words match locations you would be interested in on a general basis, it might however be a location you would not visit while on vacation. In some cases it might be that the user is not interested in visiting a location related to the interest, while an event would be a different story. Music is one example of a category where a location, like that of a general record store, might not be a place the user would like to visit, while a concert with a band he or she likes would be very interesting. The results from the RDF-model and the positive comments in the follow up support this assumption. Other examples are sport and travel. Even if you are interested in travelling, you probably won't be interested in visiting a travel agency while you are abroad. A travel event however, like an exhibition might be a different matter. Concerning the sports related interest; one tester made a specific remark about this, that sports was something she

POI: “Travelbag” - Category: TravelAgency (topic: “Travel”) (c)

POI: ”India Tourism” - Category: TravelAgency (topic: ”Travel”) (c)

POI: ”Gymbox” - Category: SportsCentre (topic: ”Sports”) (c)

Table 6.15: Examples of locations where the test participants did not want to go *on vacation*. The word listed as “topic” was the query term.

POI: “Rica Oslo Hotel” - Category: TourismHotel (topic: “Oslo”) (l)

POI: “Oslo hospital” - Category: TramStop (topic: “Oslo”) (l)

POI: “Oslo domkirke” - Category: PlaceOfWorship (topic: “Oslo”) (l)

POI: “Deichmanske biblioek, filial gamle Oslo” - Category: Library (topic: “Oslo”) (l)

POI: “Sveriges ambassad i Oslo” - Category: Embassy (topic: “Oslo”) (l)

POI: “Europcar Oslo” - Category: CarRental (topic: “Oslo”) (l)

POI: “Oslo Bymuseum” - Category: TourismMuseum (topic: “Oslo”) (l)

POI: “Oslo Mikrobryggeri” - Category: Pub (topic: “Oslo”) (l)

POI: “JET Oslo Sporveisgata” - Category: Fuel (topic: “Oslo”) (l)

POI: “Thomas Cook” - Category: BureauDeChange (topic: “Thomas”) (l)

POI: “Maria Bebudelses kirke” - Category: PlaceOfWorship (topic: “Maria”) (l)

POI: “Watches of Switzerland” - Category: Shopwatch (topic: “Switzerland”) (l)

Table 6.16: Points of interest related to proper noun topic words. The word listed as “topic” was the query term.

liked to do at home, but on vacation you would rather do something else than go to the gym. The sport example came up with different test participants where some answered “yes/maybe” and some answered “no”. At this stage the application can not differentiate between locations where some people would consider going to on vacation while others would not. Hence it is included because it is related to the user’s interest.

The follow up questions show that many of the participants have been to Oslo many times. Familiarity with locations in Oslo leads to rejection of POIs that could have been “yes/maybe’s” if the user did not know it. Furthermore the topic word “Oslo” was generated for many of the test participants resulting in points of interest with Oslo in the title. The word Oslo has no semantic meaning in this context and generated POIs as showed in Table 6.16 which are not related to the user’s interests. Other proper noun query terms also recommend POIs which does not directly relate any interest of the user. Hence proper nouns should maybe have been excluded from the query terms.

6.2.4 Test set-up

One explanation for the overall advantage to stay.com, might be the way the testing was set up. Because of the limited information available in many of the points of interest in the Semantic

Tourist, the subjects were asked to choose between “yes/maybe”, “no” and “dont know what this is” (see Section 5.2.1). In retrospect it may not have been the right choice to combine yes and maybe. Many of the test participants answered “yes/maybe” to all the hotels and all the restaurants. This could be considered in favour of stay.com, because as a tourist you are most likely to only stay at one hotel in each city. You will however most likely visit more than one restaurant, but the choice to visit usually relies on more information than what was provided in the test. Therefore these categories give an advantage to stay.com, because the Semantic Tourist does not recommend “necessity categories” like these by default.

The pilot test results indicated that it was right to adjust for a small number of input documents. However the upper limit of 40 topics might have been too high. Thinking that each topic translates to an interest, it might be that the users with large profiles has liked different entities in the same category/topic. Comparing results between single test participants show that 40 topics could be too many (in most cases). People are likely to have fewer interest. In retrospect making changes like these based on the people with small profiles in the pilot test, was not a good idea. It might be better to just accept that users with smaller profiles will get less personalization.

Looking at the results from comparing the different models of Semantic Tourist against stay.com one could wonder if there was not enough test participants to provide any significant results. The sample size could have been higher, but there were not enough time or resources to do this. However, the same data was used in all the comparisons and there where two tests with $p < .050$, showing that the sample size was big enough to produce some significant results.

To summarize; the Semantic Tourist *is* able to generate a reliable model of a user’s personal interest as shown by all the participants recognizing their Facebook profile in the topic model. For most user’s who at least use the “like” buttons as intended there seems to be enough information in Facebook to generate this model. The problem arises when trying to apply the model to the tourist domain. As discussed above there are potential for improvements in both the “label-query” and the “category-query”. We are able to generate points of interest tailored to the user. The problem is discerning between the kind the user would visit while at home, those he would visit while on vacation and those that are independent of location. A second problem is that most people also like, or even feel obligated, to visit traditional tourist attractions which usually are not related to their personal preferences.

6.3 Evaluation of the research project

One of the test sessions revealed a design issue concerning processing of a lot of user info. More investigation of the Facebook and Android SDK would prevented this from happening. The issue was not discovered in development or in pilot testing. As the test session with this participant was interrupted, it was resumed at a later time with out implication on the test.

Looking back at the testing session, it could have been a little shorter. This could have been implemented by selecting fewer points of interests to be evaluated.

The variety of test participants can be questioned. All were friends or acquaintances about the same age, and most study something related to IT. However the data collection showed that there was a wide variety of likes among the subjects. As the selection of participants was small and from a quite homogeneous group (age and educational background) of people the representative reliability of the research is not as good as it could have been (Neuman, 2011). It is also hard to determine the potential user group of the application, thus the results can not be said to representative for the population.

Chapter 7

Conclusion and future work

7.1 Conclusion

To answer the research questions raised in this thesis, a tourist application that utilize information from the user's Facebook profile to generate two models of the user's interests was created. It uses two different semantic technology approaches to create the models. This research differs from previous research because of the source of the data for the user models combined with the technologies utilized to create the models. Shapira et al. (2012) used information derived from content the user had published on Facebook as a supplement or replacement information when collaborative systems suffers from the cold-start problem. Their research indicates that information from the user's Facebook profile can significantly improve results when information is sparse. This research supports the findings in their research; that Facebook is a good source of information about the user's interests.

Both Lee et al. (2011) and Pennacchiotti and Gurumurthy (2011) used topic modelling to discern the user's interests from a social media source. Lee et al. (2011) used an additional topic model to create a time aspect to improve the recommendations base on the topic model. Their study showed high precision in their recommendations. Pennacchiotti and Gurumurthy (2011) also show good results with high recall for their topic modelling approach to find new friends for a user who share similar interests. This research also show that topic modelling can reveal acceptable semantic model of the user's interest although the recommendations based on the models were mixed. The answers from the follow up questions show that 7 out of the 10 test participants recognized the topic model created from their Facebook profile without any remarks.

The results indicate a possible problem when applying the model to the tourist domain. Even if the points of interest are tailored to the user's interests, it becomes problematic to discern between the kind that he or she would visit while on vacation, and those the he or she would only visit while at home. A second problem is that many people like or feel slightly obligated to visit at least some of the traditional tourist attractions.

The other semantic technology used in the project was RDF modelling. It was used to model the user's Facebook likes. The RDF-model was used to retrieve the names/titles of the user's activity and music related interest. These were subsequently used to find events from eventful.com that matched these interests. The results from the recommended event which this model was based on are promising. In addition test participants remark events as a useful addition to tourist application.

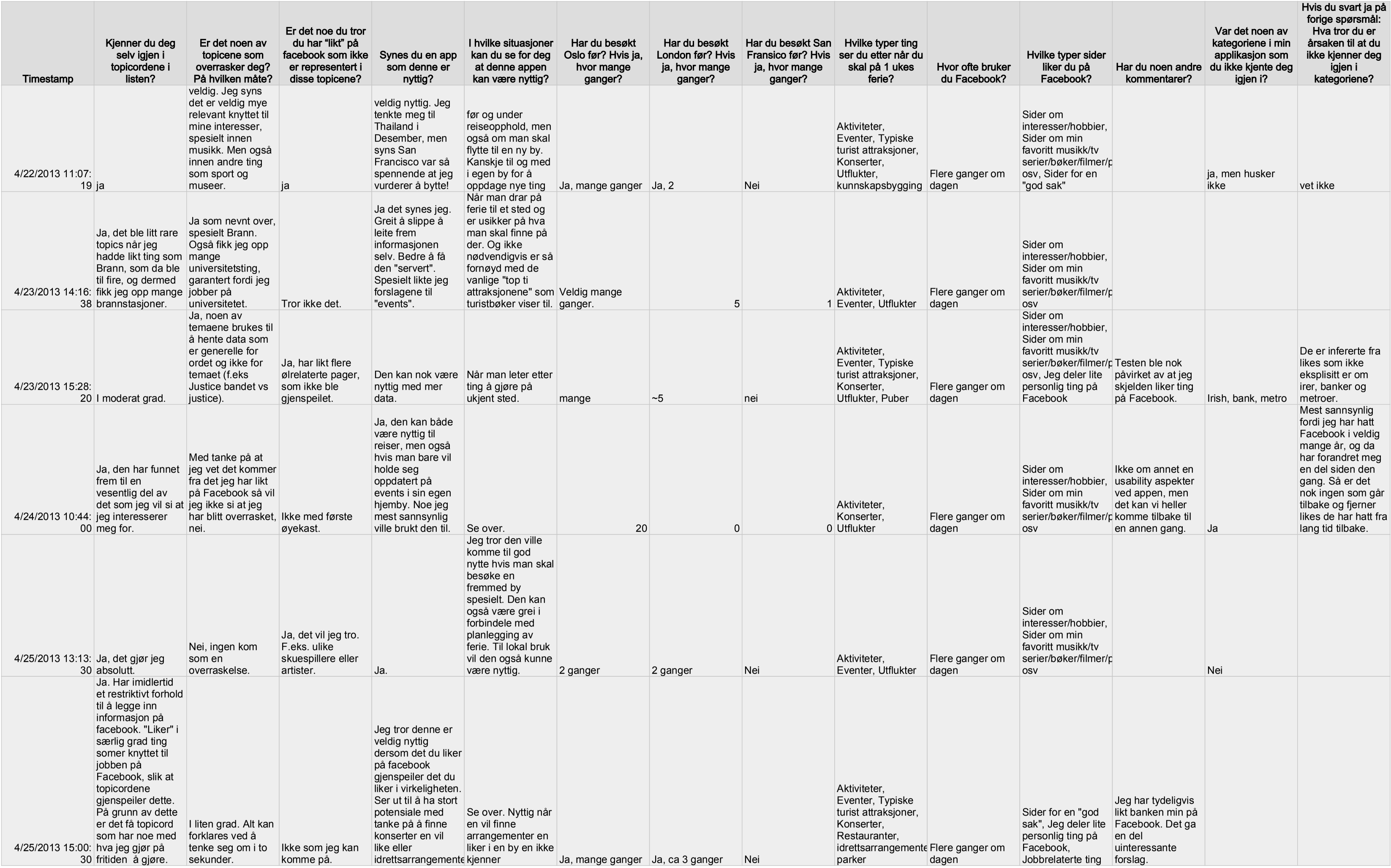
The application created in this project was successful in acquiring evaluation data that helped answer the research questions. The thesis has confirmed previous findings and produced indications of new findings. Despite the fact that the evaluation data was limited, due to the time and resource constraints of the project, the project has demonstrated that it is possible to create a reliable model of the user's preferences using semantic technologies and information collected from the user's social network. The results indicate that Facebook provide sufficient information for user's who share their preferences. The events recommended based on the RDF model of the user's likes were a hit among the test participants. While the results from the main model, the topic model, show potential for recommending points of interest in the tourist domain. There is improvement potential in both models, some of which is presented in the future work section.

7.2 Future work

Following the conclusion in this thesis, that there is a problem when applying the topic model to the tourist domain, future research should investigate the possibility of adding an ontology to semantically aid the interlinking of the topic model and third party sources of points of interest. An ontology of the topic words could help clarify the right sub category of "food" when the topic is "food-local-produce". The topics from the topic model modelled in an ontology representing the user's preferences interlinked with a domain ontology could in future work be used to aid the system in clarifying the user's preferences for on vacation activities. Adding more social network sources for the user model to increase the reliability and additional sources of points of interest are other examples of future improvements that should be considered.

Appendix A

Follow up answers

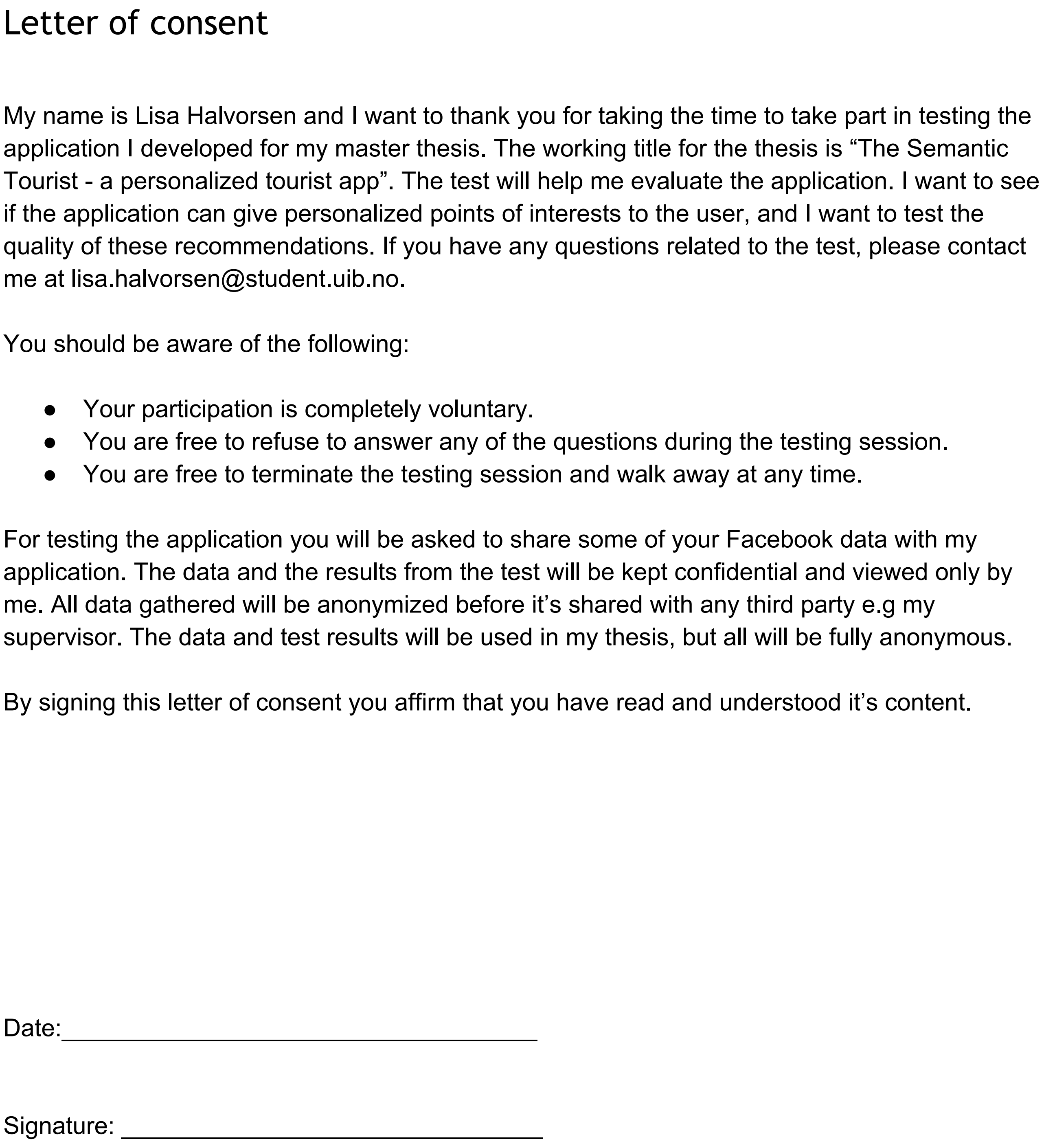


Timestamp	Kjenner du deg selv igjen i topicordene i listen?	Er det noen av topicene som overrasker deg? På hvilken måte?	Er det noe du tror du har "likt" på facebook som ikke er representert i disse topicene?	Synes du en app som denne er nyttig?	I hvilke situasjoner kan du se for deg at denne appen kan være nyttig?	Har du besøkt Oslo før? Hvis ja, hvor mange ganger?	Har du besøkt London før? Hvis ja, hvor mange ganger?	Har du besøkt San Fransico før? Hvis ja, hvor mange ganger?	Hvilke typer ting ser du etter når du skal på 1 ukes ferie?	Hvor ofte bruker du Facebook?	Hvilke typer sider liker du på Facebook?	Har du noen andre kommentarer?	Var det noen av kategoriene i min applikasjon som du ikke kjente deg igjen i?	Hvis du svart ja på forige spørsmål: Hva tror du er årsaken til at du ikke kjenner deg igjen i kategoriene?
4/22/2013 11:07:19	ja	veldig. Jeg syns det er veldig mye relevant knyttet til mine interesser, spesielt innen musikk. Men også innen andre ting som sport og museer.	ja	veldig nyttig. Jeg tenkte meg til Thailand i Desember, men syns San Francisco var så spennende at jeg vurderer å bytte!	før og under reiseopphold, men også om man skal flytte til en ny by. Kanskje til og med i egen by for å oppdage nye ting	Ja, mange ganger	Ja, 2	Nei	Aktiviteter, Eventer, Typiske turist attraksjoner, Konserter, Utflukter, kunnskapsbygging	Flere ganger om dagen	Sider om interesser/hobbier, Sider om min favoritt musikk/tv serier/bøker/filmer/p osv, Sider for en "god sak"		ja, men husker ikke	vet ikke
4/23/2013 14:16:38	Ja, det ble litt rare topics når jeg hadde likt ting som Brann, som da ble til fire, og dermed fikk jeg opp mange brannstasjoner.	Ja som nevnt over, spesielt Brann. Også fikk jeg opp mange universitetsting, garantert fordi jeg jobber på universitetet.	Tror ikke det.	Ja det synes jeg. Greit å slippe å leite frem informasjonen selv. Bedre å få den "servert". Spesielt likte jeg forslagene til "events".	Når man drar på ferie til et sted og er usikker på hva man skal finne på der. Og ikke nødvendigvis er så fornøyd med de vanlige "top ti attraksjonene" som turistbøker viser til.	Veldig mange ganger.	5	1	Aktiviteter, Eventer, Utflukter	Flere ganger om dagen	Sider om interesser/hobbier, Sider om min favoritt musikk/tv serier/bøker/filmer/p osv			
4/23/2013 15:28:20	I moderat grad.	Ja, noen av temaene brukes til å hente data som er generelle for ordet og ikke for temaet (f.eks Justice bandet vs justice).	Ja, har likt flere ølrelaterte pager, som ikke ble gjenspeilet.	Den kan nok være nyttig med mer data.	Når man leter etter ting å gjøre på ukjent sted.	mange	~5	nei	Aktiviteter, Eventer, Typiske turist attraksjoner, Konserter, Utflukter, Puber	Flere ganger om dagen	Sider om interesser/hobbier, Sider om min favoritt musikk/tv serier/bøker/filmer/p osv, Jeg deler lite personlig ting på Facebook	Testen ble nok påvirket av at jeg skjelden liker ting på Facebook.	Irish, bank, metro	De er infererte fra likes som ikke eksplisitt er om irer, banker og metroer.
4/24/2013 10:44:00	Ja, den har funnet frem til en vesentlig del av det som jeg vil si at jeg interesserer meg for.	Med tanke på at jeg vet det kommer fra det jeg har likt på Facebook så vil jeg ikke si at jeg har blitt overrasket, nei.	Ikke med første øyekast.	Ja, den kan både være nyttig til reiser, men også hvis man bare vil holde seg oppdatert på events i sin egen hjemby. Noe jeg mest sannsynlig ville brukt den til.	Se over.	20	0	0	Aktiviteter, Konserter, Utflukter	Flere ganger om dagen	Sider om interesser/hobbier, Sider om min favoritt musikk/tv serier/bøker/filmer/p osv	Ikke om annet en usability aspekter ved appen, men det kan vi heller komme tilbake til en annen gang.	Ja	Mest sannsynlig fordi jeg har hatt Facebook i veldig mange år, og da har forandret meg en del siden den gang. Så er det nok ingen som går tilbake og fjerner likes de har hatt fra lang tid tilbake.
4/25/2013 13:13:30	Ja, det gjør jeg absolutt.	Nei, ingen kom som en overraskelse.	Ja, det vil jeg tro. F.eks. ulike skuespillere eller artister.	Ja.	Jeg tror den ville komme til god nytte hvis man skal besøke en fremmed by spesielt. Den kan også være grei i forbindele med planlegging av ferie. Til lokal bruk vil den også kunne være nyttig.	2 ganger	2 ganger	Nei	Aktiviteter, Eventer, Utflukter	Flere ganger om dagen	Sider om interesser/hobbier, Sider om min favoritt musikk/tv serier/bøker/filmer/p osv		Nei	
4/25/2013 15:00:30	Ja. Har imidlertid et restriktivt forhold til å legge inn informasjon på facebook. "Liker" i særlig grad ting somer knyttet til jobben på Facebook, slik at topicordene gjenspeiler dette. På grunn av dette er det få topicord som har noe med hva jeg gjør på fritiden å gjøre.	I liten grad. Alt kan forklares ved å tenke seg om i to sekunder.	Ikke som jeg kan komme på.	Jeg tror denne er veldig nyttig dersom det du liker på facebook gjenspeiler det du liker i virkeligheten. Ser ut til å ha stort potensiale med tanke på å finne konserter en vil like eller idrettsarrangemente	Se over. Nyttig når en vil finne arrangementer en liker i en by en ikke kjenner	Ja, mange ganger	Ja, ca 3 ganger	Nei	Aktiviteter, Eventer, Typiske turist attraksjoner, Konserter, Restauranter, idrettsarrangemente parker	Flere ganger om dagen	Sider for en "god sak", Jeg deler lite personlig ting på Facebook, Jobbrelaterte ting	Jeg har tydeligvis likt banken min på Facebook. Det ga en del uinteressante forslag.		

4/25/2013 17:32:06	Kjente igjen mange / de fleste av ordene fra mine interesseområder, og det jeg har listet på facebook.	Nei igrunnen ikke, syntes de passet ganske bra	Nei	Syntes det virker veldig nyttig. Jeg fant en interessant festival jeg aldri har hørt om før ganske fort, som jeg fikk lyst å dra til i sommer!	Dersom man skal ut å reise et sted kan det være greit å se hva som skjer eller hva som er mulig å oppleve.	Ja, jeg bor i Oslo	Ja, 1	Nei	Aktiviteter, Eventer, Konserter	Flere ganger om dagen	Sider om interesser/hobbier, Sider om min favoritt musikk/tv serier/bøker/filmer/p osv, Sider for en "god sak", Usakelige ting som He-man og kamferdrops	Veldig lovendes prosjekt :)	Nei	
4/26/2013 12:06:41	Jepp!	Police station, worship,.. Skjønna ikkje korfor det er relatert til meg.	Ikkje som eg kjem på	Ja!	For maks utnytte tida man bruka i en ny by for å sjå/oppleve ting man er interessert i.	5	1	0	Aktiviteter, Eventer, Typiske turist attraksjoner, Konserter, Utflukter	Flere ganger om dagen	Sider om interesser/hobbier, Sider om min favoritt musikk/tv serier/bøker/filmer/p osv, Sider for en "god sak"		Nei	
4/26/2013 13:20:50	Ja, bra liste.	Ja, sammensetningen av topicene ser ut til å være kombinert av flere ulike likes. Eksempel Voss fjellandsby myrkdalen og hemsedal skisenter har blitt til skisenter-voss-hemsedal. Det er også flere andre topicer jeg ikke kjenner igjen (uten at jeg husker alle mine egne likes), eksempler och-bergen-för, dry-les-chamonix.	Tror jeg har liket filmen "The Rock", som henger sammen med øyen Alcatrazi San Francisco, som også er et sted jeg sikkert hadde besøkt. Dette kom ikke opp.	Ja, men litt treig lasting av POIs, mange vanlige POIs som brannstasjoner og politistasjoner.	Når man drar steder der man ikke er kjent og ikke vil bruke tid på research av området eller skaffe seg oversikt over eventer.	Ja, mange ganger	Nei	Nei	Aktiviteter, Eventer, Typiske turist attraksjoner, Konserter, Utflukter, Uteliv, bar/pub/nattklubb, restauranter	Flere ganger om dagen	Sider om interesser/hobbier, Sider om min favoritt musikk/tv serier/bøker/filmer/p osv, Sider som kommer med oppdateringer om eventer og annet. Eksempel føreforhold/åpningst i skianlegg, eller salg i butikker.	Veldig vanskelig å avgjøre interesse (avkryssing av y/n/?) kun basert på navnene i testen. Bilder og beskrivelser av steder/eventer hadde gjort det enklere.	Ja; day, Page, Garage,	Vet ikke hvor de kommer fra, kan ikke huske å ha liket noe som de kan relateres til.
4/29/2013 18:28:15	Ja, men er litt overrasket at det ikke er mer musikk-relaterte topics pga mine likes, osv.	Nei	Ja, særlig band, artister (musikk) og film/tv.	Ja, særlig eventene som poppa opp. Vanskelig å finne slike ting i turistguider, siden de ofte bare skjer en kveld, eller er ganske alternative.	Veldig nyttig ang. events, som nevnt over.	Bodd der	4	0	Eventer, Typiske turist attraksjoner, Konserter	Flere ganger om dagen	Sider om interesser/hobbier, Sider om min favoritt musikk/tv serier/bøker/filmer/p osv, Sider for en "god sak"		Nei	

Appendix B

Letter of Consent



Letter of consent

My name is Lisa Halvorsen and I want to thank you for taking the time to take part in testing the application I developed for my master thesis. The working title for the thesis is “The Semantic Tourist - a personalized tourist app”. The test will help me evaluate the application. I want to see if the application can give personalized points of interests to the user, and I want to test the quality of these recommendations. If you have any questions related to the test, please contact me at lisa.halvorsen@student.uib.no.

You should be aware of the following:

- Your participation is completely voluntary.
- You are free to refuse to answer any of the questions during the testing session.
- You are free to terminate the testing session and walk away at any time.

For testing the application you will be asked to share some of your Facebook data with my application. The data and the results from the test will be kept confidential and viewed only by me. All data gathered will be anonymized before it’s shared with any third party e.g my supervisor. The data and test results will be used in my thesis, but all will be fully anonymous.

By signing this letter of consent you affirm that you have read and understood it’s content.

Date:___________________________________

Signature: _______________________________

Appendix C

Points of interest test participants

y/n/?	Title	LDG class	Topic Word	Query
n	Metreon	Cinema	Cinema	c
n	Sundance Kabuki	Cinema	Cinema	c
n	Bing-Kong Tonc Free Masons	PlaceOfWorship	Free	l
n	Chi Sin Buddhist and Taoist Association	PlaceOfWorship	Chi	l
n	Seventh Day Adventist Japanese Church	PlaceOfWorship	Day	l
n	San Francisco Evangelical Free Church	PlaceOfWorship	Free	l
n	Central Seventh Day Adventist Church	PlaceOfWorship	Day	l
n	Philadelphian Seventh Day Adventist Church	PlaceOfWorship	Day	l
n	Seventh Day Adventist Tabernacle	PlaceOfWorship	Day	l
n	Admission Day Monument	Monument	Day	l
n	City College of San Francisco Chinatown North Beach Campus	School	North	l
n	Horace Mann Academic Middle School	School	Academic	l
n	Saint Marys Chinese Day School	School	Day	l
n	Hillwood Academic Day School	School	Academic	l
y	Radio Shack	Electronics	Radio	l
n	Cup-A-Joe	Cafe	Cup	l
n	BeBe Cleaners	DryCleaning	Dry	c
n	Billy's Dry Cleaners	DryCleaning	Dry	l
n	South Park Cleaners	DryCleaning	Dry	c
n	Eco Dry Cleaning	DryCleaning	Dry	l
n	Thick House/Golden Thread Productions	Theatre	House	l
n	Southern Police Station	Police	Station	l
n	Central Police Station	Police	Station	l
n	Union Square Garage	Parking	Garage	l
n	Golden Gateway Garage	Parking	Garage	l
n	Bed Bath & Beyond	Housewares	House	c
n	UCSF/Mission Bay	RailwayStation	Station	c

n	King & 4th	RailwayStation	Station	c
n	Van Ness St Muni	RailwayStation	Station	c
n	San Francisco Caltrain	RailwayStation	Station	c
n	Montgomery St BART/Muni	RailwayStation	Station	c
n	16th St. Mission BART	RailwayStation	Station	c
n	Brannan & The Embarcadero Muni	RailwayStation	Station	c
n	4th & King	RailwayStation	Station	c
n	Embarcadero BART/Muni	RailwayStation	Station	c
n	24th St Mission BART	RailwayStation	Station	c
n	North East Medical Services (NEMS)	Hospital	North	l
n	San Francisco Fire Department Main Office	FireStation	Station	c
n	San Francisco Fire Station 7	FireStation	Station	l
n	Battalion 2	FireStation	Station	c
n	San Francisco Fire Department Station 2	FireStation	Station	l
n	Bluxome 8 San Francisco Fire Department	FireStation	Station	c
n	San Francisco Fire Station 11	FireStation	Station	l
n	San Francisco Fire Station 41	FireStation	Station	l
n	San Francisco Fire Station 37	FireStation	Station	l
n	Fire Station Number Two	FireStation	Station	l
n	San Francisco Fire Department Station 1	FireStation	Station	l
n	Java House Breakfast and Lunch	Restaurant	House	l
n	North Beach Restaurant	Restaurant	North	l
n	Thai House Express	Restaurant	House	l
y	North Beach Museum	TourismMuseum	North	l
n	Bamboo Reef Scuba Supply	Outdoor	Outdoor	c
n	REI	Outdoor	Outdoor	c
y	Johnny Foley's Irish House	Pub	House	l
n	House of Shields	Pub	House	l
n	Pour House	Pub	House	l
y	Rogue Ales Public House	Pub	House	l
n	Chinatown Station Post Office	PostOffice	Station	l
n	Mission Station San Francisco Post Office	PostOffice	Station	l
n	Steiner Street Station San Francisco Post Office	PostOffice	Station	l
n	Sutter Station	PostOffice	Station	l
n	Macy's Station	PostOffice	Station	l
n	Pine Street Station San Francisco Post Office	PostOffice	Station	l
n	Gateway Station San Francisco Post Office	PostOffice	Station	l
n	Noe Valley Station San Francisco Post Office	PostOffice	Station	l
n	Shell Gas Station	Fuel	Station	l
n	76 Gas Station	Fuel	Station	l
n	Droubi Team Colwell Banker	Shop	Team	l
n	United States Customs House	PublicBuilding	House	l
n	Coin Op'd Wash Dry	Laundry	Dry	l

n	Super Dry Cleaners	Laundry	Dry	l
n	Professional Dry Cleaners	Laundry	Dry	l
n	123 Wash and Dry	Laundry	Dry	l

Table C.1: San Francisco points of interest from test participant with below 20% precision. Full version of Table 6.4

Appendix D

Points of interest test participants

y/n/?	Title	LDG class	Topic Word	Query
y	Clooney's Pub	Bar	Pub	l
n	Bing-Kong Tonc Free Masons	PlaceOfWorship	Free	l
y	Golden Gate Community Church of the Nazarene	PlaceOfWorship	Community	l
y	Chinese Community Church	PlaceOfWorship	Community	l
y	Community Baptist Church	PlaceOfWorship	Community	l
y	Seventh Day Adventist Japanese Church	PlaceOfWorship	Japanese	l
y	San Francisco Evangelical Free Church	PlaceOfWorship	Free	l
y	Mission Branch Library	Library	Library	c
y	Far West Library for Educational Research and Development	Library	Library	c
y	Mechanic's Institute Library and Chess Room	Library	Library	c
y	Western Addition Branch Library	Library	Library	c
y	Erik Erikson Library	Library	Library	c
y	Golden Gate Valley Branch Library	Library	Library	c
y	Presidio Branch Library	Library	Library	c
y	William E Colby Memorial Library	Library	Library	c
y	Potrero Branch Library	Library	Library	c
y	San Francisco Public Library Mission Bay Branch	Library	Library	c
n	University of California Hastings College of the Law	School	University	l
y	San Francisco Conservatory of Music	School	Music	l
y	Heald Business College	School	Business	l
n	Tenderloin Community School	School	Community	l
y	Academy of Art University	School	University	l
y	Golden Gate University - San Francisco	School	University	l

n	Dental School of the University of the Pacific	School	University	l
n	San Francisco University High School	School	University	l
y	University of California Extension Center	School	University	l
y	Music and Art Institute	School	Music	l
y	John Foley's Dueling Piano Bar	Nightclub	Piano	l
y	Cup-A-Joe	Cafe	Cup	l
n	SoMa Community Recreation Center	SportsCentre	Community	l
n	Vietnamese Community Center of San Francisco	CommunityCentre	Community	c
y	Great American Music Hall	Theatre	Music	l
n	San Francisco Community Convalescent Hospital	Hospital	Community	l
n	Potrero Hill Health Center	Hospital	Health	l
n	Mission Neighborhood Health Center	Hospital	Health	l
y	San Francisco Pier 48	FerryTerminal	Pier	l
y	Northeast Community Federal Credit Union, Tenderloin Branch	Bank	Community	l
y	Carmen's Pier 40 Restaurant	Restaurant	Pier	l
y	Music Exchange	Shopmusical_instrument	Music	l
y	California Society of Pioneers Museum and Library	TourismMuseum	Library	l
y	Mission Community Recreation Center	Park	Community	l
y	Noe's Bar	Pub	Pub	c
y	First Crush Wine Bar	Pub	Pub	c
y	Ha-Ra	Pub	Pub	c
y	Cigar Bar	Pub	Pub	c
y	San Francisco Brewing Company	Pub	Pub	c
y	Hemlock	Pub	Pub	c
y	Geary Club	Pub	Pub	c
y	Rogue Ales Public House	Pub	Pub	c
y	Kilowatt	Pub	Pub	c
y	14 Romolo	Pub	Pub	c
y	Ella's Health Spa Hot Tubs	Shop	Health	l
y	Noe Valley Music	Shop	Music	l
y	Clarion Music Center	Music	Music	c
y	Hall of Justice	PublicBuilding	Pub	c
?	United States Customs House	PublicBuilding	Pub	c
y	The Gladstone Institutes	University		
	University	c		
y	UCSF Mission Bay	University	University	c
y	University of Phoenix	University	University	c
y	Pier 52	Pier	Pier	c
?	Community Thrift Store	Shopthift_store	Community	l

Table D.1: San Francisco points of interest from test participant with above 70% precision. Full version of Table 6.2

Bibliography

PA Aek. Semantic Web Personalization. *Citeseer*, 2005. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.6070&rep=rep1&type=pdf.

Android Developers. Activities — Android Developers, 2013a. URL http://developer.android.com/guide/components/activities.html.

Android Developers. Fragments — Android Developers, 2013b. URL http://developer.android.com/guide/components/fragments.html.

Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, 21(3):372–4, March 2010. ISSN 1467-9280. doi: 10.1177/0956797609360756. URL http://www.ncbi.nlm.nih.gov/pubmed/20424071.

Matteo Baldoni, Cristina Baroglio, and Nicola Henze. Personalization for the Semantic Web. *World Wide Web Internet And Web Information Systems*, 506779(506779):173–212, 2005.

Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL Web Ontology Language Reference, 2004. URL http://www.w3.org/TR/owl-ref/#sameAs-def.

Walter Bender. Twenty years of personalization: All about the "daily me". (October), 2002.

Tim Berners-Lee. Linked Data - Design Issues, 2006. URL http://www.w3.org/DesignIssues/LinkedData.html.

Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009. ISSN 15526283. doi: 10.4018/jswis.2009081901. URL http://www.citeulike.org/user/omunoz/article/5008761.

David M Blei. Introduction to Probabilistic Topic Models. pages 1–16, 2011.

DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. URL `http://dl.acm.org/citation.cfm?id=944937`.

Jan Blom. Personalization: a taxonomy. *CHI'00 extended abstracts on Human factors in ...*, (April):1–2, 2000. URL `http://dl.acm.org/citation.cfm?id=633483`.

A. Bryman. *Social research methods*. Oxford University Press, Incorporated, 2008. ISBN 9780199202959. URL `http://books.google.no/books?id=O7a2QAAACAAJ`.

Robin Burke. Knowledge-based recommender systems. *Encyclopedia of Library and Information Science: ...*, pages 1–23, 2000. URL `http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Knowledge-based+recommender+systems#0`.

Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002. ISSN 0924-1868. URL `http://dx.doi.org/10.1023/A:1021240730564http://www.springerlink.com/index/N881136032U8K111.pdf`.

D Carmel, N Zwerdling, and I Guy. Personalized social search based on the user's social network. *Proceedings of the 18th ...*, 2009. URL `http://dl.acm.org/citation.cfm?id=1646109`.

Dbpedia.org. wiki.dbpedia.org : About, 2008. URL `http://dbpedia.org/About`.

Zakaria Elberrichi, Abdelattif Rahmoun, and MA Bentaalah. Using WordNet for text categorization. *The International Arab Journal ...*, 5(1):16–24, 2008. URL `http://ccis2k.org/iajit/PDF/vol.5,no.1/3-37.pdf`.

Facebook. Documentation, 2013a. URL `https://developers.facebook.com/docs/`.

Facebook. Facebook Investors, 2013b. URL `http://investor.fb.com/releasedetail.cfm?ReleaseID=761090`.

Facebook. Create a page, 2013c. URL `https://www.facebook.com/pages/create/?ref_type=sitefooter`. `https://www.facebook.com/pages/create/?ref_type=sitefooter`.

Facebook. Facebook's Growth In The Past Year — Facebook, 2013d. URL `https://www.facebook.com/photo.php?fbid=10151908376831729&set=a.10151908376636729.1073741825.20531316728&type=1&theater`.

Facebook. Facebook APIs, 2013e. URL `https://developers.facebook.com/docs/reference/apis/`.

Facebook. Facebook developers - android change log 3.x, 2013f. URL `https://developers.facebook.com/android/change-log-3.x/`.

Facebook.com. Welcome to Facebook — Log in, sign up or learn more, 2013. URL `https://www.facebook.com/`.

Haiyan Fan and MS Poole. What is personalization? Perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and ...*, (May 2013):37–41, 2006. URL `http://www.tandfonline.com/doi/full/10.1080/10919392.2006.9681199`.

The Apache Software Foundation. Apache Lucene - Apache Solr, 2012. URL `http://lucene.apache.org/solr/`.

Angel García-Crespo, Javier Chamizo, Ismael Rivera, Myriam Mencke, Ricardo Colomo-Palacios, and Juan Miguel Gómez-Berbís. SPETA: Social pervasive e-Tourism advisor. *Telematics and Informatics*, 26(3):306–315, August 2009. ISSN 07365853. doi: 10.1016/j.tele.2008.11.008. URL `http://linkinghub.elsevier.com/retrieve/pii/S0736585308000683`.

Serge Garlatti and Sébastien Iksal. Declarative Specifications for Adaptive Hypermedia Based on a Semantic Web Approach. pages 81–85, 2003.

TR Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, (April), 1993. URL `http://secs.ceas.uc.edu/~mazlack/ECE.716.Sp2011/Semantic.Web.Ontology.Papers/Gruber.93a.pdf`.

David Hall and Saikat Kanjilal. Latent Dirichlet Allocation - Apache Mahout - Apache Software Foundation, 2013. URL `https://cwiki.apache.org/confluence/display/MAHOUT/Latent+Dirichlet+Allocation`.

John Hebeler, Matthew Fisher, Ryan Blace, Andrew Perez-Lopez, and Mike Dean. *Semantic Web Programming*. John Wiley \& Sons Inc., 2009. ISBN 978-0-470-41801-7.

Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *Mis Quarterly*, 28(1):75–105, 2004. URL `http://www.wirtschaftsinformatik.de/pdf/wi2006_2_133-142.pdf`.

Andreas Hotho, Alexander Maedche, and Steffen Staab. Ontology-based Text Document Clustering. pages 1–13, 1998.

Katerina Kabassi. Personalizing recommendations for tourists. *Telematics and Informatics*, 27(1):51–66, February 2010. ISSN 07365853. doi: 10.1016/j.tele.2009.05.003. URL `http://linkinghub.elsevier.com/retrieve/pii/S073658530900029X`.

Leslie Lamport. LaTeX – A document preparation system, 2010. URL `http://www.latex-project.org/`.

SS Lee, T Chung, and D McLeod. Dynamic item recommendation by topic modeling for social networks. *Information Technology: New . . .*, 2011. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5945352`.

LinkedGeoData.org. linkedgeodata.org : About, 2013. URL `http://linkedgeodata.org/About`.

Mark Little, Savas Parastatidis, Ian Robinson, Gregor Roth, Brian Sletten, Stefan Tikov, Steve Vinoski, and Jim Webber. InfoQ Explores REST. 2010.

Apache Mahout. Apache Mahout: Scalable machine learning and data mining, 2011. URL `http://mahout.apache.org/`.

David Maltz and K Ehrlich. Pointing the way: active collaborative filtering. *Proceedings of the SIGCHI conference on Human . . .*, (May), 1995. URL `http://dl.acm.org/citation.cfm?id=223930`.

Adriana M. Manago, Michael B. Graham, Patricia M. Greenfield, and Goldie Salimkhan. Self-presentation and gender on MySpace. *Journal of Applied Developmental Psychology*, 29(6): 446–458, November 2008. ISSN 01933973. doi: 10.1016/j.appdev.2008.07.001. URL `http://linkinghub.elsevier.com/retrieve/pii/S0193397308000749`.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. An Introduction to Information Retrieval, 2009. URL `http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf`.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Matthew Michelson and SA Macskassy. Discovering users' topics of interest on twitter: a first look. *Proceedings of the fourth workshop on . . .*, pages 73–79, 2010. URL `http://dl.acm.org/citation.cfm?id=1871852`.

Stuart E. Middleton, Nigel R. Shadbolt, and David C. De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22(1):54–88, January 2004. ISSN 10468188. doi: 10.1145/963770.963773. URL http://eprints.ecs.soton.ac.uk/8926/1/tois2004.pdfhttp://portal.acm.org/citation.cfm?doid=963770.963773.

Miquel Montaner. A Taxonomy of Recommender Agents on the Internet. pages 285–330, 2003.

W.L. Neuman. *Social Research Methods: Qualitative and Quantitative Approaches.* MyResearchKit Series. Pearson Higher Ed USA, 2011. ISBN 9780205786831.

J.F. Nunamaker Jr and M. Chen. Systems development in information systems research. In *System Sciences, 1990., Proceedings of the Twenty-Third Annual Hawaii International Conference on*, volume 3, pages 631–640. IEEE, 1990. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=205401.

J.F. Nunamaker Jr, M. Chen, and Titus D.M. Purdin. Systems development in information systems research. *Journal of Managemental Information Systems*, 7:89–106, 1990. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=205401.

Communications Of and T H E Acm. Personalization A U SER -C ENTERED D ESIGN A PPROACH to. 43(8), 2000.

Alice Oh. Topic models applied to online news and reviews - google tech talk august 11, 2010. http://www.youtube.com/watch?v=1wcX4fEdNUo, 2010. URL http://www.youtube.com/watch?v=1wcX4fEdNUo.

OpenStreetMap.org. OpenStreetMap Wiki, 2013. URL http://wiki.openstreetmap.org/wiki/Main_Page.

Evan W Patton and Deborah L Mcguinness. The Mobile Wine Agent : Pairing Wine with the Social Semantic Web. *World*, 2009.

Marco Pennacchiotti and Siva Gurumurthy. Investigating topic models for social media user recommendation. *Proceedings of the 20th international conference companion on World wide web - WWW '11*, page 101, 2011. doi: 10.1145/1963192.1963244. URL http://portal.acm.org/citation.cfm?doid=1963192.1963244.

Peter Pirolli and Stuart Card. Information foraging in information access environments. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '95*, pages

51–58, 1995. doi: 10.1145/223904.223911. URL http://portal.acm.org/citation.cfm?doid=223904.223911.

FOAF project. The Friend of a Friend (FOAF) project — FOAF project. URL http://www.foaf-project.org/.

Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF, 2008. URL http://www.w3.org/TR/rdf-sparql-query/#QueryForms.

Elaine Rich and Computer Sciences. Users are individuals : - individualizing user models. (May 1981), 1983.

Howard J Seltman. Experimental Design and Analysis. 2012.

Bracha Shapira, Lior Rokach, and Shirley Freilikhman. Facebook single and cross domain data for recommendation systems. *User Modeling and User-Adapted Interaction*, 23(2-3): 211–247, September 2012. ISSN 0924-1868. doi: 10.1007/s11257-012-9128-x. URL http://link.springer.com/10.1007/s11257-012-9128-x.

Rogers Y. Preece J. Sharp, H. *SInteraction Design: Beyond Human-Computer Interaction.* Wiley, 2007. ISBN 0470018666. URL http://www.id-book.com/.

James Shore. James Shore: The Art of Agile Development: Spike Solutions, 2007. URL http://jamesshore.com/Agile-Book/spike_solutions.html.

SpringSource.org. Spring Tool Suite — SpringSource.org, 2013. URL http://www.springsource.org/sts.

The Apache Software Foundation. Apache Maven, 2013a. URL http://maven.apache.org/.

The Apache Software Foundation. Apache Jena - Apache Jena, 2013b. URL http://jena.apache.org/.

Leman Pınar Tosun. Motives for Facebook use and expressing "true self" on the Internet. *Computers in Human Behavior*, 28(4):1510–1517, July 2012. ISSN 07475632. doi: 10.1016/j.chb.2012.03.018. URL http://linkinghub.elsevier.com/retrieve/pii/S0747563212000842.

Brendon Towle and Clark Quinn. Knowledge based recommender systems using explicit user models. *. . . of the AAAI Workshop on Knowledge-Based Electronic . . .*, pages 74–77, 2000. URL http://www.aaai.org/Papers/Workshops/2000/WS-00-04/WS00-04-011.pdf.

Simine Vazire and Samuel D Gosling. e-Perceptions: personality impressions based on personal websites. *Journal of personality and social psychology*, 87(1):123–32, July 2004. ISSN 0022-3514. doi: 10.1037/0022-3514.87.1.123. URL http://www.ncbi.nlm.nih.gov/pubmed/15250797.

W3C. RDF Vocabulary Description Language 1.0: RDF Schema, 2004. URL http://www.w3.org/TR/rdf-schema/#ch_label.

W3C. Ontologies - W3C, 2013. URL http://www.w3.org/standards/semanticweb/ontology.