Thesis Design

Roy Hoeymans
Master Information Sciences
VU University, Amsterdam

Internal supervisor: Victor de Boer
External supervisor: Timo Kouwenhoven
Company: DNV-GL

Effective recommendation in a knowledge base
The SKYbrare case study

1. Introduction

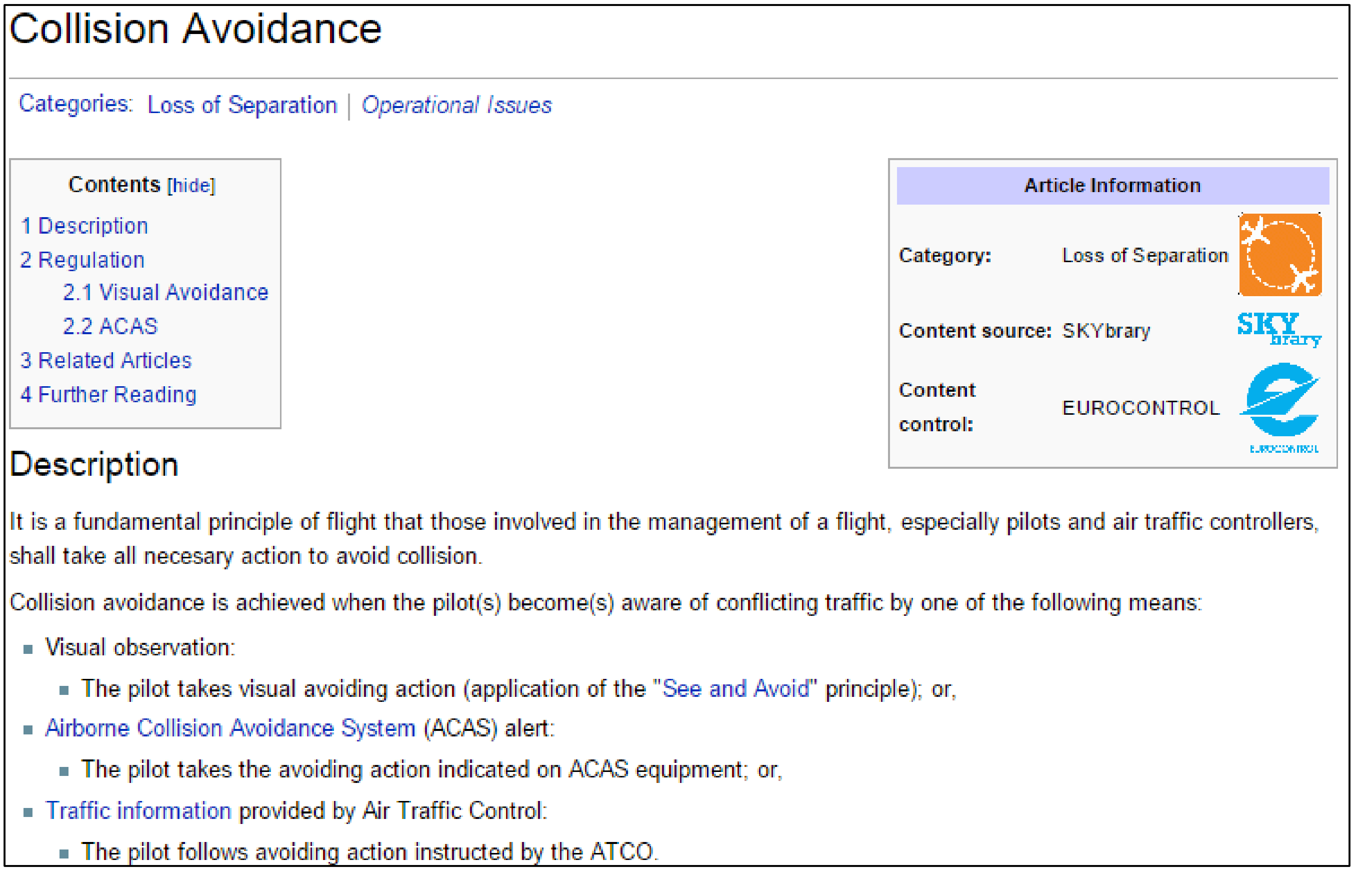
This document describes the thesis design for my master project. The master project is external and will be executed at DNV-GL. In this project I will investigate the possibility of a recommender system in relatively small knowledge bases that contain highly specialized information. I will be using SKYbrary, a knowledge base on aviation and air traffic management safety built by DNV-GL as a case study. A detailed problem description will be addressed in section 2, section 3 explains the relevance of the problem based on a review of the related literature, section 4 describes the resulting research question, section 5 explains the research methodology that will be used to answer this research question, and section 6 shows the planning of the project.

2. Problem description

SKYbrary is an internet platform containing a large amount of aviation and air traffic management safety knowledge that was launched by EUROCONTROL, the European Organisation for the Safety of Air Navigation, and built by DNV-GL, a classification company that mainly operates in the aviation, ship transport, energy, oil and gas, and information technology sector. The objective of SKYbrary is to be a single point of access to knowledge on these topics, and it is set up as a Wiki site. Since its conception in 2008, it has grown into a knowledge base that is internationally recognized and supported by the most important aviation organizations, institutions, and regulators.

Unlike Wikipedia, not every user can add or edit content on the website. The content is written by a small group of aviation safety experts and each article is evaluated carefully. Today, SKYbrary (www.skybrary.aero) contains around 5000 articles, and each week a featured article is sent to its 20.000 subscribers.

To make it easier for users to navigate through the large amount of knowledge that SKYbrary has collected, DNV-GL wants to find out if it is possible to include a selection of recommended articles in the 'Article Information'-box in the top right corner of each article (Fig. 1.) by using a certain recommender system technique or algorithm.



Collision Avoidance

Categories: Loss of Separation | *Operational Issues*

Article Information	
Category:	Loss of Separation
Content source:	SKYbrary
Content control:	EUROCONTROL

Description

It is a fundamental principle of flight that those involved in the management of a flight, especially pilots and air traffic controllers, shall take all necesary action to avoid collision.

Collision avoidance is achieved when the pilot(s) become(s) aware of conflicting traffic by one of the following means:

- Visual observation:
  - The pilot takes visual avoiding action (application of the "See and Avoid" principle); or,
- Airborne Collision Avoidance System (ACAS) alert:
  - The pilot takes the avoiding action indicated on ACAS equipment; or,
- Traffic information provided by Air Traffic Control:
  - The pilot follows avoiding action instructed by the ATCO.

Fig. 1. Screenshot of a SKYbrary article

Recommender systems are a hot topic for large websites, like amazon.com and youtube.com. They use data about the item that the user is currently viewing or the user's browsing behaviour to generate links that this user might also be interested in. However, websites like Amazon have an immense amount of data available. SKYbrary is relatively small and contains very specialized knowledge.

The challenge of this research is to investigate if it is useful and feasible to implement a recommender system using a relatively small amount of data, by experimenting with several types of recommendation algorithms.

3. Related literature

3.1. About recommender systems
Recommender systems are software tools and techniques that provide (often personalized) recommendations or suggestions about what the user of a website should do next. [1] Examples of this are Youtube recommending which video's to watch, Amazon recommending which items you might want to buy, and Netflix suggesting movies that you might like to watch. This way users can cope with the huge amount of alternatives or products that are available.

With recommender systems playing an increasingly important role in popular websites like the ones mentioned earlier, more and more research articles on this topic are being published. [2] In a paper by Park (2011) 164 articles on recommender systems were analysed and ordered by field, e.g. books, movies, and music. They show that by far most of these articles are in the fields of movies and shopping, possibly because of its commercial value; recommender systems in those fields can increase the chance of users buying more products.

Since SKYbrary is a non-profit knowledge base, we are not interested in the commercial value of recommender systems. What we ultimately want to achieve with the implementation of such a system is an increase in user satisfaction [1, p.5]. If the recommendations are relevant and interesting, the user can navigate through the site more easily, making him more likely to stay on the site longer, revisit the site, and find the site more useful.

Recommender systems produce these suggestions based on a certain algorithm or technique. The literature mentions several types of these. The techniques that are the most relevant are summarized briefly in the following sub-sections:

3.1.1. Data mining

Data mining algorithms have a large impact on the recommender systems field and their use can be found in any recommender system approach. [1, p.39, 8] Typically, the data is cleansed and filtered first so that it can be used by machine learning techniques. Each object in the data has certain attributes, and a computed distance measure compared to other objects. An unsupervised or supervised classification algorithm then determines which objects are nearest to each other and recommends those to the user.

3.1.2. Content-based filtering

A content-based recommender system makes recommendations by analysing the content of an item and finding similarities with other items. A prerequisite for this approach to work is the availability of rich metadata that describes the content. [9]

3.1.3. Link-based filtering

Another way of finding recommended or related pages is the use of the hyperlink structure of the web. If a page *v* contains a link to page *w*, it is likely that page *w* is related. [5] [6]

3.1.4. Collaborative filtering

Collaborative filtering is a common approach in recommender systems where suggestions are made by comparing a user’s behaviour to the behaviour of other users that have similar characteristics. This personalized recommender system technique is often used in the field of movies and television, but also in the prediction of the next webpage that a user might want to visit. [10]

3.2. Current limitations

Several of these approaches have been used in research papers, as is described in a literature survey by Beel [3]. What is remarkable is that the combined research efforts can not provide a clear answer to which technique is the most useful for a recommender system. In some cases content-based filtering performed better than collaborative filtering or the other way around. According to Beel, the ambiguity of these results is mainly caused by malpractices, such as not properly or not at all evaluating the recommendations produced by the system, or not including a baseline. [3]

3.3. Novelty

In the context of SKYbrary, a content-based approach based on metadata or a link-based approach based on hyperlinks in an article seems the most promising, because the content data and metadata is easy to acquire from a knowledge base that is set up as a Wiki. However, DNV-GL has also gathered statistics over the last 7 years about user navigation behaviour via Google Analytics, like from which page within SKYbrary visitors of a certain article came from, or which pages were viewed in each visitor’s session. I think it would also be interesting to investigate the effectiveness of an algorithm based on user navigation behaviour, which I will refer to as user navigation-based. Because a large percentage of the users on SKYbrary, and Wiki sites in general, are new visitors and users who don’t have an account, detailed data about specific user characteristics and preferences is not largely available. Thus, I will also investigate the effectiveness of an algorithm that is based on anonymous user navigation data, which I have not found in the current literature on recommender systems.

I will also include a thorough evaluation of the algorithms. I will use the expertise of the content editors of SKYbrary to verify if the algorithms produce sensible recommendations. The evaluation method is described in more detail in subsection 5.4.

4. Research question

Main research question:

- What is an effective algorithm or combination of algorithms for a relatively small digital knowledge base containing highly specialized information?

To answer the main research question, an answer must be provided for the following subquestions:

1. Which type of algorithm works best for a recommender system for such a knowledge base according to the content editors of SKYbrary?
2. How do the recommendations made by algorithms differ from the manually selected ‘Related Links’-section of each article?

5. Research methodology

5.1. Data
DNV-GL has gathered Google Analytics statistics about SKYbrary over the last 7 years. Data about how the users navigated through the website, for example from which page within SKYbrary visitors of a certain article came from, is especially useful. This data can be used for the user navigation-approach. For a content-based approach, the SKYbrary database and the SKYbrary article content and metadata is useful. All the content is stored in the database, and the Wiki software creates the actual pages, meaning that the HTML code of each SKYbrary article is structured in the same way. For example, each article contains a <div>-element with id “#catlinks” containing links to categories that the article is part of, and a <span>-element with id “#Related_Articles” with links to related articles.

5.2. Approach
First I will extract and pre-process the data from Google Analytics and SKYbrary so that it can be used as input for the algorithms I want to test. I will test three different algorithms: content-based, user navigation, and a hybrid version of those. To build these algorithms I will use LensKit, an open-source tool written in Java for building and evaluating recommender systems. This tool contains pre-built user-based and item-based algorithms that I will modify and configure for the SKYbrary data.

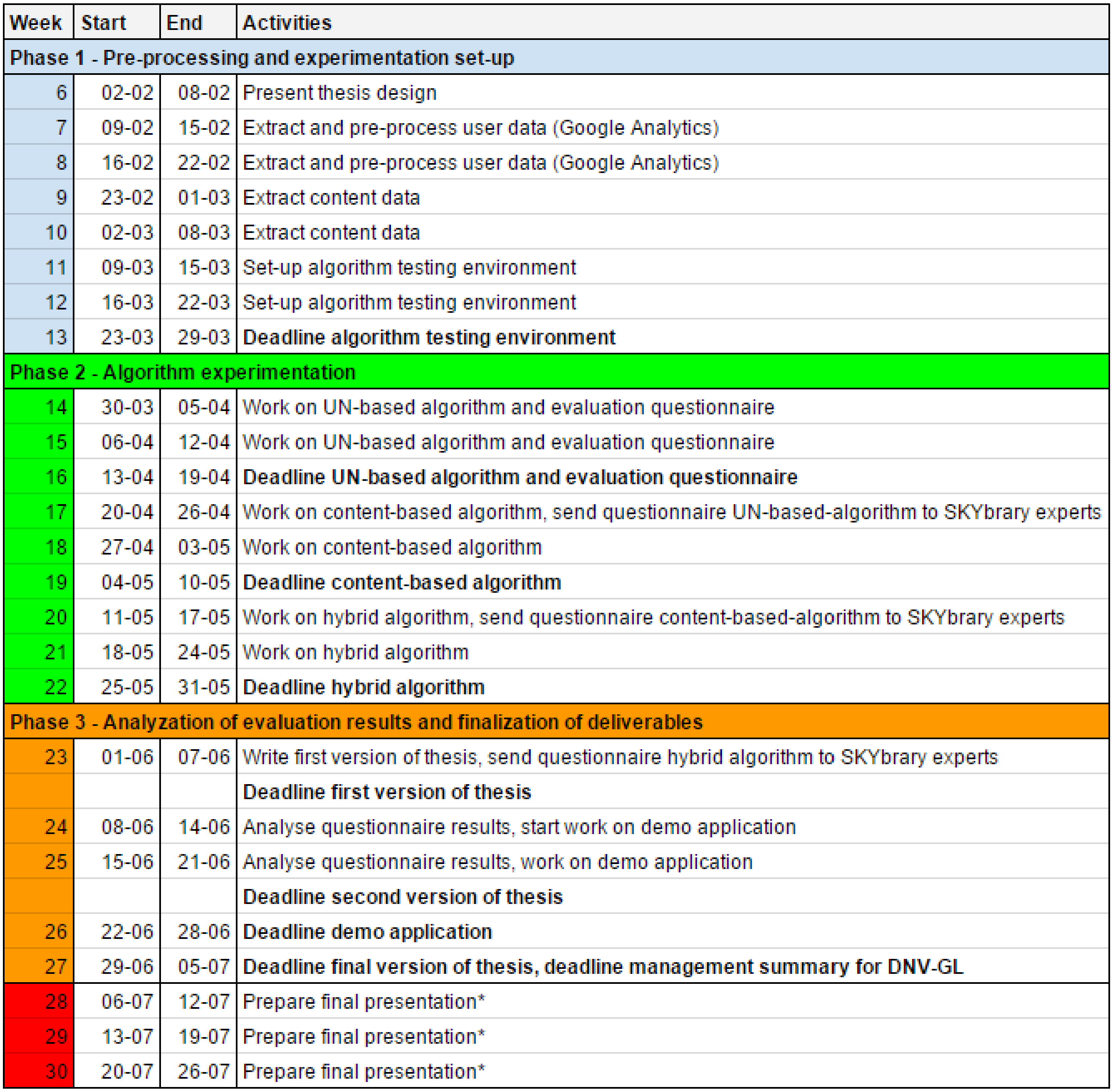
5.3. Evaluation
What I will include in the approach is an evaluation of the quality of the recommendations given, which is often neglected in the research in this field. [3] The resulting recommendations that an algorithm comes up with can be verified by the experts on aviation safety that create and edit the content on SKYbrary. The way I want to do this is by creating a questionnaire that shows the content editors of SKYbrary an article, plus the recommendations that the algorithm produces for it. They can then rate each recommended article on a Likert scale from 1 to 5 (not at all relevant to very relevant). This way I can analyse which of these algorithms work best according to the experts of SKYbrary, which allows me to answer subquestion 1.

5.4 Result
The result of this project will consists of the following:

- An evaluation of several recommender system algorithms, including an advice for DNV-GL about which one would work best for SKYbrary, as well as a more general advice for websites similar to SKYbrary.
- A demo application that shows the resulting recommendations for a selected SKYbrary article and algorithm.

The exact planning of the research methodology is presented in the next section.

6. Planning and deadlines



Week	Start	End	Activities
Phase 1 - Pre-processing and experimentation set-up			
6	02-02	08-02	Present thesis design
7	09-02	15-02	Extract and pre-process user data (Google Analytics)
8	16-02	22-02	Extract and pre-process user data (Google Analytics)
9	23-02	01-03	Extract content data
10	02-03	08-03	Extract content data
11	09-03	15-03	Set-up algorithm testing environment
12	16-03	22-03	Set-up algorithm testing environment
13	23-03	29-03	**Deadline algorithm testing environment**
Phase 2 - Algorithm experimentation			
14	30-03	05-04	Work on UN-based algorithm and evaluation questionnaire
15	06-04	12-04	Work on UN-based algorithm and evaluation questionnaire
16	13-04	19-04	**Deadline UN-based algorithm and evaluation questionnaire**
17	20-04	26-04	Work on content-based algorithm, send questionnaire UN-based-algorithm to SKYbrary experts
18	27-04	03-05	Work on content-based algorithm
19	04-05	10-05	**Deadline content-based algorithm**
20	11-05	17-05	Work on hybrid algorithm, send questionnaire content-based-algorithm to SKYbrary experts
21	18-05	24-05	Work on hybrid algorithm
22	25-05	31-05	**Deadline hybrid algorithm**
Phase 3 - Analyzation of evaluation results and finalization of deliverables			
23	01-06	07-06	Write first version of thesis, send questionnaire hybrid algorithm to SKYbrary experts
			Deadline first version of thesis
24	08-06	14-06	Analyse questionnaire results, start work on demo application
25	15-06	21-06	Analyse questionnaire results, work on demo application
			Deadline second version of thesis
26	22-06	28-06	**Deadline demo application**
27	29-06	05-07	**Deadline final version of thesis, deadline management summary for DNV-GL**
28	06-07	12-07	Prepare final presentation*
29	13-07	19-07	Prepare final presentation*
30	20-07	26-07	Prepare final presentation*

** The date of my final presentation is not yet known*

The project plan is divided into the following phases:

Phase 1 - Pre-processing and experimentation set-up
The first phase will be used to extract and pre-process the user navigation data collected in Google Analytics and the content of SKYbrary so that it can be used as input for algorithms. I will also try some existing recommender system tools and select on that can be used to set-up an algorithm testing environment.

Phase 2 - Algorithm experimentation
In this phase, the user-behaviour based, the content-based and the hybrid algorithms will be developed in the testing environment. After each algorithm is completed, I will send a questionnaire to the SKYbrary experts to test if the recommendations it produces are relevant.

Phase 3 - Analyzation of evaluation results and finalization of deliverables
In the final phase I will process and analyse the results of the evaluations, as well as finalize the deliverables that are the result of this project: the master thesis, the management summary for DNV-GL, and the demo application.

References

[1] Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P.B. (2011) “Recommender Systems Handbook”

[2] Deuk, H. P.; Hyea, K. K.; Il, Y. C.; Jae, K. K. (2012) “A Review and Classification of Recommender Systems Research”, Deuk Hee Park, Expert Systems with Applications, vol. 39, issue 11, 1 september 2012, pp.10059 - 10072.

[3] Beel, J.; Langer, S. (2013) “Research Paper Recommender Systems: A Literature Survey”, RepSys '13 Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, pp. 15-22.

[4] Adomavicius, G.; Tuzhilin, A.; (2005) “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions”, IEEE Transactions on Knowledge and Data Engineering, vol. 17, issue 6.

[5] Dean, J.; Henzinger, M. R. (1999) “Finding Related Pages in the World Wide Web”, WWW '99 Proceedings of the eighth international conference on World Wide.

[6] Chirita, P. A.; Olmedilla, D.; Nejdl, W. (2004) “Finding Related Pages Using the Link Structure of the WWW”, L3S and University of Hannover, Germany.

[7] Adomavicius, G.; Mobasher, B.; Ricci, F.; Tuzhilin, A (2011) "Context-Aware Recommender Systems." AI Magazine, vol. 32, no. 3, 2011.

[8] Schafer, J. (2009) “The Application of Data-Mining to Recommender Systems”, Encyclopedia of data warehousing and mining.

[9] Rajaraman, A.; Ullman, J. D. (2011) "Data Mining". *Mining of Massive Datasets*. pp. 1–17.

[10] Al Murtadha, Y.; Sulaiman, M. N.; Mustapha, N.; Udzir, N.I. (2011) “Improved web page recommender system based on web usage mining,” in Proceedings of the 3rd International Conference on Computing and Informatics (ICOCI), 2011, pp. 8–9.