

University of Bristol  
Computer Science Department  
Merchant Venturers Building,  
Woodland Road  
BS8 1UB, Bristol, UK



Plan  
“Weather Talk”,  
extracting weather information by text  
mining

by

**Student:** Vasileios Lampos [v17342]

**Supervisor:** Professor Nello Cristianini

**Marker1:** Dr James A. R. Marshall

**Marker2:** Dr Walterio W. Mayol-Cuevas

**COMSM2100**

“Project Specification and Design, Advanced”

Faculty of Engineering  
Department of Computer Science

May 2008

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Aim and Objectives</b>	<b>2</b>
2.1	Aim of “Weather talk” . . . . .	2
2.2	Objectives of “Weather talk” . . . . .	3
<b>3</b>	<b>Planning each objective and progress made</b>	<b>5</b>
3.1	Objectives and progress made . . . . .	5
3.1.1	Assure resources used as input data . . . . .	5
3.1.2	Assure resources for grounding truth . . . . .	6
3.1.3	Fetch the resources into a database . . . . .	6
3.1.4	Decide the weather states of the model . . . . .	7
3.1.5	Build a weather ontology . . . . .	7
3.1.6	Classification of the observable data . . . . .	7
3.1.7	First level evaluation . . . . .	7
3.1.8	Data fusion and re-evaluation . . . . .	7
3.1.9	Application of a Hidden Markov Model and re-evaluation . . . . .	8
3.1.10	Build a weather map . . . . .	8
3.1.11	Expand the model in order to infer more weather information* . . . . .	8
3.1.12	Apply the model to other data types by redesigning* . . . . .	8
3.1.13	Write dissertation and prepare for the presentation . . . . .	9
3.2	Risk analysis . . . . .	9
<b>4</b>	<b>Timetable</b>	<b>11</b>
4.1	Complementary information about the timetable . . . . .	11
	<b>Bibliography</b>	<b>14</b>

# Chapter 1

## Introduction

It will become obvious that the initial plan of the proposal has been changed. The main reason was that “working” on the project in terms of finding ways to combine knowledge derived from the background research, pointed to this direction. The content of the plan is distributed as follows; in chapter 2 we state the aim of this project together with a list of required or optional objectives which we want to achieve. In chapter 3 we analyze the concept of each objective, we provide a description on how we plan to achieve it, and we refer to the progress made to some of them. Then, we provide a risk analysis for all the stages of the project. Finally, chapter 4 presents the exact timetable of “Weather talk”.

## Chapter 2

# Aim and Objectives

In this chapter we set the broader aim of this project as well as the more specific objectives which break the whole procedure into well-defined and achievable parts.

### 2.1 Aim of “Weather talk”

A very general and unrealistic<sup>1</sup> aim of “Weather talk” is the investigation and implementation of a system capable of making conclusions by combining random web information. A more realistic aim of this project is to design a model which then will be implemented in order to extract weather information from observable web data. Weather was chosen for the fact that our conclusions can be evaluated (availability of ground truth).

Being more specific about what we try to accomplish, we will “observe” and collect public information from the web, which is unrelated with official weather data, and then we will build a model in order to decide the weather state (or the major weather condition) of a location for a specific day. We will build this model in steps, adding each time a new feature, which on the one hand increases the total complexity of the system, but on the other we hope that improves the final decision’s accuracy.

---

<sup>1</sup> in terms of the short period of time that is available for an MSc project.

## 2.2 Objectives of “Weather talk”

We would like to have independent, stand-alone objectives in order to reduce the general risk in this project, but usually the combinations of knowledge are of high interest, and thus we could not avoid a significant degree of coupling<sup>2</sup> between the several objectives.

A list of the **initial** objectives of “Weather talk” is:

1. Collect or find ways of dynamic collection of the data that will be used for this project (observable web information).
2. Find official weather resources which will be used for evaluating our decisions.
3. Implement an application capable to fetch the resources (from [1] and [2]) and store them into a database.
4. Decide for the weather states used in the model based on [2].
5. Build a weather ontology which conforms with the needs of the project based on [4].
6. Design and implement classification on the collected data by using Bayesian inversion (based on [3] and [4]).
7. Evaluate the model based on results of [2] and [6].
8. Expand the model by using data fusion and perform re-evaluation (based on [1] - [7]).<sup>3</sup>

A list of the **optional**, but with *high probability* to meet, objectives is:

9. Design and implement a Hidden Markov Model (hereinafter HMM) in order to improve the final conclusions (based on [6] - [8]).
10. Implement a graphical representation of the inferred weather states.<sup>4</sup>

A list of our **optional**, but with *low probability* to meet,<sup>5</sup> objectives is:

11. Expand the model in order to include approximations on the values of temperature and humidity in addition to the already inferred weather state.

---

<sup>2</sup> *Coupling* is a term used in object oriented programming languages in order to indicate whether a unit of a program is linked with other units. A big number of links leads to “tight” coupling, which we would like to avoid.

<sup>3</sup> Objective [7] uses all the previous steps since it will perform on a different data source.

<sup>4</sup> This objective can be implemented separately by mapping hypothetical weather states, and as a result it is independent.

<sup>5</sup> These objectives depend on the previous work and we will try to accomplish them only if the general progress of the project points to that direction.

12. Try to apply the model in other types of web observable information, such as opinions about products or individuals. This will demand to redesign most the parts of the model.

Finally, a “**default**” objective for this project is:

13. Based on the results of [1] - [10], write the dissertation text and prepare the presentation.

By completing objectives [1] - [8], we will have succeeded in the main aim of the project. Objective [9] introduces a further challenge, but depends on the success of the previous objectives; otherwise its application is impossible. On the other hand, objective [10] does not depend on the success of the model, but consists of design issues which do not bring any research interest. Objectives [11], and [12] are listed and included in the whole plan, in order to indicate how we are going to continue our research in case of very early and unexpected success; we believe that these objectives will be left out for future research with high probability.

## Chapter 3

# Planning each objective and progress made

In this chapter we make a more extended reference to each objective mentioned in the second chapter, including any progress which has already been made. We also make clear how objectives interact with each other. In the end, we provide a risk assessment plan for this project.

### 3.1 Objectives and progress made

We will follow the numbering provided in chapter 2 as far as objectives are concerned. A set of small sections dedicated to each one of the objectives follows. Sections' titles followed by an asterisk (\*) describe optional objectives with very small probability of achieving them.

#### 3.1.1 Assure resources used as input data

As we have mentioned in the review of the project, there are a lot of methods to implement crawling and fetch observable information with a stand-alone mechanism. On the other hand, for the needs of this project we need a very strong crawler, able to fetch as many data of interest as possible (*e.g.* search engine crawler). Consequently, we will fetch information directly from search queries which map to articles<sup>1</sup> using well-known, dedicated, and massive tools such as Google Blog Search,<sup>2</sup> Google News,<sup>3</sup> and Technorati.<sup>4</sup> We have ended up with

---

<sup>1</sup> Most of the times the search result provides a feed with the contents of every result (*i.e.* the articles of interest).

<sup>2</sup> Google Blog Search, <http://blogsearch.google.com/>.

<sup>3</sup> Google News, <http://news.google.com/>.

<sup>4</sup> Technorati, <http://technorati.com/>.

two solutions on this, which will be followed in parallel at the beginning. After receiving the first experimental results (objective [7]) we may only use the most effective approach.<sup>5</sup> The approaches are:

- Collect all weather related search results for the United Kingdom (hereinafter UK) for a specific day, and then see the locations mentioned in them, which can be different every day. A minimal example, using Google News is, a search query like “*weather OR sunny OR rainy OR cloudy location:uk*”.<sup>6</sup>
- Collect all weather related search results for specific locations in the UK (for a specific day). A minimal example, using Google Blog Search, is a search query like “*Bristol weather OR sunny OR rainy OR cloudy*”.<sup>7</sup>

### 3.1.2 Assure resources for grounding truth

As it was explained in the project review, we will have to base the design of our model to the form of the official weather feeds. This is the main reason that completing this objective defines the next one, as well as the completion of this objective depends on the demands of the next objective. We will collect, official weather observations for the major locations in the UK. Observations update themselves frequently (every half or one hour); we will collect all the available observations for a location during a day, and then decide for the major weather state (a mapping off all the official observations into one weather state which will be called “*official weather state*”). The decision process will be implemented in objective [5]. Possible resources will be BBC weather feeds,<sup>8</sup> or Yahoo! weather feeds.<sup>9</sup>

### 3.1.3 Fetch the resources into a database

This objective refers to the implementation of a program able to fetch web data and store them in a database. More specifically, we want to fetch the contents of the results of the search queries of [1] into a database in order to be able to parse them for elements of interest afterwards. We could have placed the implementation of this module, right after completing the design of the weather ontology for the reason that the ontology will be used in order to

<sup>5</sup> Of course, as the model is being extended, effectiveness may not be stable.

<sup>6</sup> Information is missing from the query. For a more detailed preview, try [http://news.google.com/news?as\\_q=&svnum=10&as\\_scoring=r&ned=uk&btnG=Google+Search&as\\_epq=&as\\_oq=weather+sunny+rainy+cloudy&as\\_eq=&as\\_qdr=&as\\_drrb=b&as\\_mind=2&as\\_minm=5&as\\_maxd=3&as\\_maxm=5&as\\_nsrc=&as\\_nloc=UK&as\\_occt=any&aq=f](http://news.google.com/news?as_q=&svnum=10&as_scoring=r&ned=uk&btnG=Google+Search&as_epq=&as_oq=weather+sunny+rainy+cloudy&as_eq=&as_qdr=&as_drrb=b&as_mind=2&as_minm=5&as_maxd=3&as_maxm=5&as_nsrc=&as_nloc=UK&as_occt=any&aq=f).

<sup>7</sup> Information is missing from the query. For a more detailed preview, try [http://blogsearch.google.com/blogsearch?as\\_q=Bristol&num=10&hl=en&ctz=-60&c2coff=1&btnG=Search+Blogs&as\\_epq=&as\\_oq=weather+sunny+rainy+cloudy&as\\_eq=&bl\\_pt=&bl\\_bt=&bl\\_url=&bl\\_auth=&as\\_qdr=a&as\\_drrb=b&as\\_mind=2&as\\_minm=5&as\\_miny=2008&as\\_maxd=3&as\\_maxm=5&as\\_maxy=2008&lr=&safe=off](http://blogsearch.google.com/blogsearch?as_q=Bristol&num=10&hl=en&ctz=-60&c2coff=1&btnG=Search+Blogs&as_epq=&as_oq=weather+sunny+rainy+cloudy&as_eq=&bl_pt=&bl_bt=&bl_url=&bl_auth=&as_qdr=a&as_drrb=b&as_mind=2&as_minm=5&as_miny=2008&as_maxd=3&as_maxm=5&as_maxy=2008&lr=&safe=off).

<sup>8</sup> BBC weather feeds, <http://backstage.bbc.co.uk/data/WeatherFeeds>.

<sup>9</sup> Yahoo! weather feeds, <http://developer.yahoo.com/weather/>.

make queries to the search engines (mentioned in objective [1]). However, this module is generic, and we want to start working on it as early as possible.

### 3.1.4 Decide the weather states of the model

The weather states will be decided by the content of the resources found in [2]. The major weather states highlighted in official weather observations, will be the base on building the weather ontology in the next step.

### 3.1.5 Build a weather ontology

After defining the set of weather states for “Weather talk”, a weather ontology will be built based on them. This ontology might not follow the general concepts of an ontology, as for the purposes of this project we would like to have descriptions for the weather states, and not a general semantic description of the weather. The first step, in completing this objective, will be to write an eXtensive Markup Language Schema (hereinafter XML Schema), which will describe the structure of our XML file, which initially will hold the ontology (we may build a more formal ontology using Protégé).<sup>10</sup> We will investigate ways of using WordNet to extend our ontology, as it is described in [NP03], [KMV00], and [KL02].

### 3.1.6 Classification of the observable data

This objective consists of a design and an implementation part. A customized to the needs of the project version of Naïve Bayesian classification will be applied to the retrieved information. We have based our model on the general theory about Bayesian inversion on [Mit97] and [Liu07].<sup>11</sup> The final outcome of this step would be a decision about the major weather state of several locations for a specific day.

### 3.1.7 First level evaluation

This objective has a design and an implementation part as well. We will evaluate the results of objective [6], as described in chapter 5 of the review. If the average percentage of success is small, we will have to apply changes to the previous steps, and especially to [6], and [5].

<sup>10</sup> Protégé ontology editor, <http://protege.stanford.edu/>.

<sup>11</sup> The model is described in chapter 3 of the review.

### 3.1.8 Data fusion and re-evaluation

Again, this objective has a design and an implementation part and depends on the successful completion of [7]. We will pick another context of observable information related to weather, such as traffic information, and we will create mappings to weather states. Using the module implemented in [3], we will fetch the related information into our database. Then, all the previous steps will be applied in order to extract a weather state from the new observable information. Finally, we will merge the prediction made in [6] with this one (data fusion). Evaluation is going to be applied using the same method as in [7].

### 3.1.9 Application of a Hidden Markov Model and re-evaluation

In the review, we have already made a description (and a discussion) about how we are going to use Hidden Markov Model. In case this approach does not meet the demands of the project or does not offer the intended results, we will have to return to the previous objectives and increase the accuracy of the final result for each stage ([6], [7], [8]). The success of this objective is based on our hypothesis, that the weather states between two consecutive days are highly coupled. If this is proven to be a wrong assumption, then we may skip this objective.

### 3.1.10 Build a weather map

In order to present the functionality of our system in a more interactive and attractive way, we will create a map by interpolating between the inferred locations (for which we have decided a major weather state). Google Maps API will be used for this objective.<sup>12</sup> This objective does not add any scientific research in our project, and as a result it will be carried out, only if time allows.

### 3.1.11 Expand the model in order to infer more weather information\*

For this task, we will expand the “Weather talk” model in order to infer more information about the weather such as temperature and humidity estimations. Apart from being optional, this is a totally experimental, with very small chances of success goal for the reasons which have been mentioned in chapter 5 of the review. However, it is a point of interest; the next objective has a more interesting research feature, and depending on the availability of time, we may choose to work on [12], rather than [11].

---

<sup>12</sup> Google Maps Application Program Interface documentation, <http://code.google.com/apis/maps/documentation/>.

### 3.1.12 Apply the model to other data types by redesigning\*

This an optional, but interesting objective with very limited chances of achieving, not only because it overloads the project work, but also due to the nature of the problem which tries to address. The main concept of this objective is to use the base of the previous model, in order to build a more sophisticated one; the new model will be able to extract opinions about products, individuals, locations, etc. We may use naïve Bayes classification for text using a vocabulary, as it is described in [MN98] or try a different way to identify and classify sentiments in text (again using a dictionary), as in [KH04]. Then, we may apply an HMM extension to our design.

### 3.1.13 Write dissertation and prepare for the presentation

The dissertation will be written using LaTeX typesetting system.<sup>13</sup> For the purposes of the interim report, we have “created”<sup>14</sup> a template which will also be used for the dissertation. We have scheduled time slots during the summer, in order to complete parts of the dissertation, when a milestone is achieved. Most of the writing will be carried out during September.

## 3.2 Risk analysis

In this section, we have included a minimal risk assessment for “Weather talk” presented in Table 3.1.<sup>15</sup> For each task there is a Probability and Severity factor able to take values from 1 to 10. The total score of a risk is their product ( $min = 1$ ,  $max = 100$ ). We have also listed some actions that may resolve the situation.

---

<sup>13</sup>LaTeX project, <http://www.latex-project.org/>.

<sup>14</sup> Double quotes were used as we did not create from a scratch a template; we have just combined several templates on the web (which were not bound with a copyright) with our own previous LaTeX material.

<sup>15</sup> Table’s structure and the general concept of this section were taken from “Planning, Risk, and Reflection” slides of Andrew Charlesworth, <http://www.cs.bris.ac.uk/Teaching/advanced/project/howtowrite/HowToResearch-PlanningRiskandReflection.pdf>.

TABLE 3.1: Risk analysis of “Weather talk”. Probability and Severity can take values from 1 to 10. Score is equal to (Probability  $\times$  Severity).

<b>Risk</b>	<b>Probability</b>	<b>Severity</b>	<b>Score</b>	<b>Actions</b>
Official weather resource unavailable	<i>1</i>	<i>9</i>	<i>9</i>	<i>Have multiple resources, store data in a database for later use.</i>
Unsuccessful first level evaluation	<i>2</i>	<i>10</i>	<i>20</i>	<i>Apply changes to the weather ontology. If the results are not improved, change the set of weather states. If we still do not get the expected results change Bayesian inversion model.</i>
Unsuccessful evaluation of objective <i>i</i>	<i>3</i>	<i>7</i>	<i>21</i>	<i>Change the design of objective <i>i</i>. If the results are not improved, then try the solutions of unsuccessful first level evaluation.</i>
Equipment failure	<i>2</i>	<i>10</i>	<i>20</i>	<i>Store all the work in an online database. Use departmental or university’s computer facilities.</i>
Illness or personal misfortune	<i>3</i>	<i>3</i>	<i>9</i>	<i>We have already assigned bigger time slots in each objective in order to deal with such situations.</i>

# Chapter 4

## Timetable

In this section we present the timetable of the project followed by some additional notes in order to specify the time lines more clearly.

### 4.1 Complementary information about the timetable

The timetable of “Weather talk” (Figure 4.1) consists of thirteen main tasks. “(D)”, and “(I)” at the end of a task denote whether we refer to the design or the implementation part of the task respectively. A list with the exact time lines as well as with some additional comments on each objective follows:<sup>1</sup>

- 1) From *10/06* to *14/06*. We have already made a 20% progress on this.
- 2) From *12/06* to *16/06*. We will do this in parallel with [1], as the main concept for both is searching. A progress of 25% has already been made.
- 3) From *15/06* to *23/06*. We hope that by the *15/06* we would have finished searching for resources. In any case, we will start implementing the data collection module; we can still search for resources in parallel.
- 4) From *24/06* to *27/06*. This is an important step; however, the four days we plan to dedicate on it may be a lot.
- 5) From *26/06* to *02/07*. This is a key step of this project. If needed, we can “borrow” time from [4].
- 6.1) From *03/07* to *07/07*. We have already made a 10% progress on this.

---

<sup>1</sup> From 01/01 to 04/01, means that we plan to work on something for four days (01, 02, 03, 04) of the 1<sup>st</sup> month (January).

- 
- 6.2) From 06/07 to 14/07. This is the first major implementation objective.
  - 7.1) From 15/07 to 17/07. We have already made a 15% progress on this.
  - 7.2) From 15/07 to 21/07. This is the first *milestone* of “Weather talk” as it indicates whether the previous work is on the right way. If the evaluation of the model is not successful, then we will have to go back and redesign some parts of the scheme.
  - 8.1) From 22/07 to 26/07. We have already made a progress of 5% on this objective.
  - 8.2) From 24/07 to 03/08. This is the second *milestone*.
  - 9.1) From 04/08 to 07/08. We have already made a progress of 10% on this objective.
  - 9.2) From 05/08 to 18/08. This is the third *milestone*.
  - 10) From 17/08 to 28/08. If the progress of the project allows, we may start implementing this earlier than scheduled.
  - 11) (*optional*) From 29/08 to 04/09. Overall progress will indicate whether we are going to pursuit this objective.
  - 12) (*optional*) From 30/08 to 06/09. Overall progress will indicate whether we are going to pursuit this objective.
  - 13) From 20/07 to 21/07, 02/08 to 03/08, 17/08 to 18/08, and 05/09 to 29/09. This is the final *milestone*. We have scheduled to write some parts of the report during the summer, each time we complete a milestone of the project.

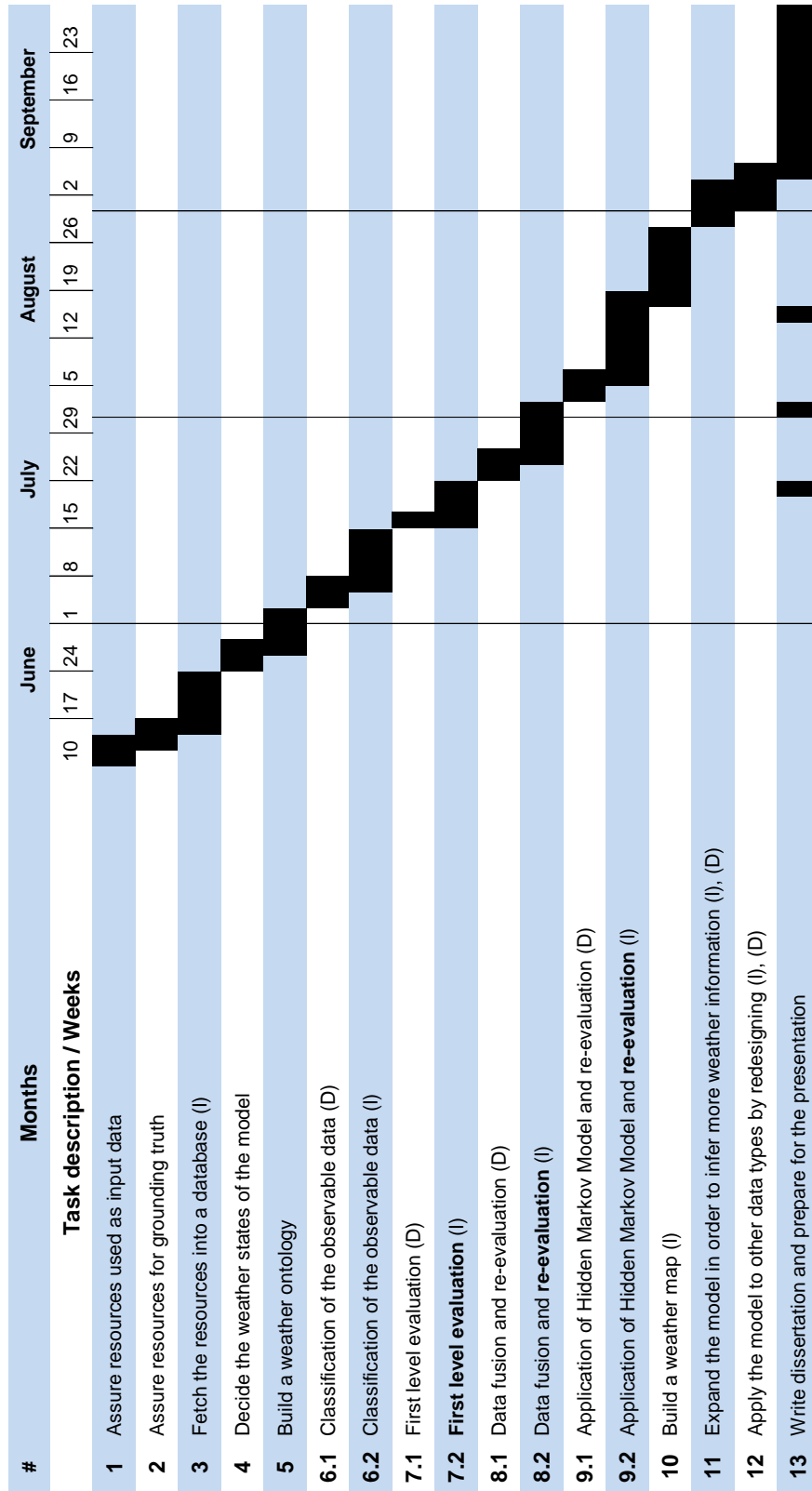


FIGURE 4.1: “Weather talk” - Timetable.

# Bibliography

- [KH04] S.M. Kim and E. Hovy. Determining the sentiment of opinions. *Proceedings of COLING*, 4:1367–1373, 2004.
- [KL02] L. Khan and F. Luo. Ontology construction for information selection. *Tools with Artificial Intelligence, 2002.(ICTAI 2002). Proceedings. 14th IEEE International Conference on*, pages 122–127, 2002.
- [KMV00] J.U. Kietz, A. Maedche, and R. Volz. A method for semi-automatic ontology acquisition from a corporate intranet. *EKAW-2000 Workshop Ontologies and Text, Juan-Les-Pins, France, October 2000*, 2000.
- [Liu07] B. Liu. *Web data mining: exploring hyperlinks, contents, and usage data*, chapter 3, pages 87–97. Springer, 2007.
- [Mit97] T.M. Mitchell. *Machine Learning. WCB*, chapter 6, pages 154–200. Mac Graw Hill, 1997.
- [MN98] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*, 752, 1998.
- [NP03] I. Niles and A. Pease. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages 412–416, 2003.