etBlogAnalysis - Analysing tourism Weblogs and forums using statistical and computer linguistic methods for quality control

Klemens Waldhör
Krems Research Forschungsgesellschaft mbH
klemens.waldhoer@kremsresearch.at

Abstract

Blogs and forums related to travel and tourism have become quite popular in the last years. Travellers use them to share their experiences with other travellers and the whole Web community. Content produced in this way can give an invaluable support in developing and improving tourism products. The number of travel blogs, forums and entries has exploded in recent years. This leads to the uncomfortable situation for tourism managers and marketers that they cannot read and check all the entries potentially mentioning their enterprise, organisation or destination. Semiautomatic routines – software applications, Web crawlers – can be of great help to check the Web automatically based on some criteria defined by the user and create summary reports, which further on can be used to react to entries or use them for marketing purposes. This paper describes the approach used in the etBlogAnalysis project, which aims for developing a combined crawler / sentiment extraction application for the tourism domain. It combines a simple and robust linguistic parsing methodology with information and terminology extraction methods in order to determine relevant utterances on expression (statement) level. The expressions are modelled based on their polarity and their power. The resulting valued expressions are aggregated on various levels (segment, paragraph, document) and overall reports are produced thereof. A simple application produces warnings for an enterprise (e.g. hotel) if too many negative entries have been written about it.

Keywords: Tourism, Weblog, Blog, Travel Forum, Computational Linguistics, Sentiment Analysis, Virtual Communities, Linguistic Modelling, Text Classification, Text Understanding.

Virtual Communities

Virtual Communities (VC) have become very popular over the past years. Blogs make a very important contribution to VCs but they are by far not the only one. eBay, (evaluation) forums, YouTube, Wikis, 2nd Life or Google’s Panoramio are examples of virtual communities, which enable people to share their experiences with others all around the world. While communities and services like YouTube or 2ndLife focus on the visual experiences blogs, travel wikis or forums concentrate on the written word, which by no means excludes photos or videos. Nevertheless, the communication focus is “text”. Text means that information is easily accessible, indexable and searchable. Search engines have no problems indexing entries in blogs or forums. Blogs are searched using specialised search engines, e.g. http://blogsearch.google.com, http://technorati.com. As a result, content provided by tourists can be easily found by everyone - taken into account that due to the huge number of matches the relevant content may not appear on the first page of the search results.
This leads to some essential problems for the tourism industry. Blogs have originally evolved from online diaries to document travel experiences to allow communication with relatives at home instantly and keep them up-to-date. The number of Weblogs and forums has exploded in the last years although research shows that in Virtual Communities especially in forums only a few participants are responsible for most of the entries.

A key issue in tourism marketing is to stay on top of the current tourist behaviour developments. What does the tourist expect, what does he/she think about a destination, where will he/she travel in the next years, how does he/she communicate his/her experiences, and what does he/she state about a specific hotel? While in the pre-Web time his/her utterances were quite limited in range - just personally communicated to relatives and friends – Weblogs spread opinions fast and make them accessible all over the Internet world. Virtual communities allow a person to make his/her experiences – either positive or negative ones - available world wide without much effort. They are just a mouse click away. A negative evaluation of a hotel may immediately result in fewer bookings for the hotel. A very positive description about the friendliness of the servers in a restaurant and about excellent food served may increase the number of guests drastically. And to make the situation worse, there is no real way – except good service – to influence this. Although tourism organisations ask their members to react on (negative or false) utterances in blogs and request deletion of those blog entries, it is questionable if this makes much sense given the fact that once entries are out, they can be easily archived (www.archive.org) and thus also found. Furthermore the impression of censorship can lead to further negative coverage.

Questions for tourism organisations

Not long ago, tourism organisation (TO) focused their Web presence on a website or webportal, beautiful photos, search engine optimisation etc... The organisation acted as an active contributor to the Web while the consumer had to behave “passive”, reading the information he/she found in the Web. As both blogs and forums now provide a simple way to move from the passive consumer position into an active content provider position, TOs need to draw their attention also to VCs. In comparison to a private “homepage”, an entry in a Blog or forum has much more impact”, if a traveller shares his experiences. VCs are in most cases operated by organisations, which optimise their content with regard to the different search engines. In addition, features like feeds, trackbacks, cross linking, etc. make those Web sites much more powerful tools.

Nevertheless there are crucial questions TOs have to consider, e.g. the problem of trust. Whom do consumers trust more? Do they trust the TO or the entry of a blogger in the Web?

This question is obviously not easy to answer. The following figure summarises some of the questions, which TOs have to consider when they want to make use of blogs or forums.

- How do those reports influence readers = potential guests?
  - How do people react to positive / negative statements in blogs and Web forums?
  - How do guests react when they realise a gap between report and reality?
    - Whom will guests blame? TO? Author?
  - Who should be blamed for „negative evaluations and reports“?
    - How should a TO react to negative statements?
Figure 1: Questions for TOs

Classification of uses cases for blog analysis

Blogs and forums are not a coherent mode of communication. Both writers and reader may have different goals. A traveller will read a blog entry with a different view compared to a hotel manager. In the remainder of the paper the focus is given to tourism organisations (destinations, hotels, travel agencies, etc.), which are interested in understanding and learning from the entries written by travellers and guests.

Thus, it is intended to understand C2C blogs: Obviously, B2B and B2C blogs may also contain valuable information. The authors intend to develop an approach, which focuses on C2C blogs. During the course of the discussion it will be shown that this restriction is not too hard and allows getting valuable insights into the guest community. This is the first classification we look at.

A second classification asks if the TO analyses blogs, which are written about the TO itself (e.g. a hotel wants to know what guests write about the hotel) or if the TO is interested in texts written about other TOs (e.g. a TO being interested in the performance of another destination).

A third classification deals with the goal of the analysis:

- Quality of service analysis and quality of service improvements by checking positive and negative utterances of guests, SWOT analysis;
- Marketing: using blogs for marketing purposes, e.g. identifying testimonials, which could be used in marketing campaigns, which key words are associated with a destination;
- **Trend research**, e.g. by detecting frequently visited tourism offers and their features. This deals also with the anticipation of trends and the shift of interest in the bloggers community, especially with trend setters in the community.

A **fourth classification** deals with the owner (provider) of a blog: Is the blog operated by an independent organisation or by the tourist organisation itself? This has great influence on the credibility and trust of the reader. A reader may per se assume that a blog operated by a TO will have some bias towards a positive presentation. On the other hand, “independent providers of blog and forums” may also get into bad light if they provide paid links for hotel bookings.

This relates to the **fifth classification** asking who the **author of the blog entries** is. Are the blog or forum entries written by the travellers – the ones who consume the tourism product - or are they written by people being part of a TO? This may also heavily influence the reception of the content by the reader. The problem here is that TOs may use the role of the “traveller” to add “positive evaluations”. This is quite a problematic behaviour as it generally will drive blogs and forums into a negative direction as being not trustworthy. In some cases, readers can easily determine that entries have not been written by real tourists if this can be inferred by the language – e.g. writing positive comments in German about a Turkish hotel where the reader immediately recognises that the author’s mother tongue is not German.

A **sixth classification** is the tourism sector, e.g. a blog about transportation, accommodation, travel agency, etc.

**Manual versus automatic blogs analysis**

A key problem for the participants in the travel value chain is now to identify all the relevant utterances done by bloggers and forum writers all over the world. As the social Web rapidly expands day by day, this is not a really easy task, especially when one has to check the community sites manually. Depending on the intention of the analysis, several hundreds or thousands of entries have to be read in regular intervals. Obviously, this is a task one cannot run on a daily, weekly or monthly basis. Thus, most organisations will tend to perform such an analysis just at certain occasions. But even then the amount of time needed is considerable.

We start with a short discussion about the advantages of using software-based support analysis application before presenting the etBlogsAnalysis:

As explained before, the number of blogs and forums in the Web has exploded. A manual inspection of the blogs and the entries is nearly impossible. Entries are added somewhere every second. Consequently, inspecting Web logs is an ongoing job, which has to be done at regular intervals. While for human beings this job gets quite boring, software agents are perfect in running repetitive jobs over and over. In addition, it is easy to run a software agent again with slightly modified parameters.

Human inspection tends to focus on certain aspects of a text and may miss important facts. And humans are not really good with repetitive tasks. After having read 200 entries, one gets tired and bored and may tend to ignore important facts of the text. In addition, humans tend to overestimate the occurrence of specific items based on previous experiences, thus, also resulting in biased selections. While this might also be true for software agents as they depend
on the settings provided by the user, they will not bias their results depending on the number of entries they have to check.

Software agents will store the retrieved content in an adequate manner (e.g. in a database). They will avoid looking at duplicates over and over again.

In the end it is also a cost question. A good analysis will require several days a month and the costs can easily go up to several thousands Euros. Software will be less expensive, at least in the long run.

**The etBlogsAnalysis Project**

In this chapter we explain the approach which has been chosen by the etBlogsAnalysis project, which is a sub project of the ANET (Austrian Network for eTourism) project etNewProduct.

The main goal of the etBlogsAnalysis project is to provide a simple and efficient software research prototype allowing handling the following problem areas:

a) **Classification of blog and forums entries with regard to different categories.** This results in a multidimensional classification, e.g. identifying an entry as dealing with Upper Austria (the region), hotel (accommodation), and climbing (sports).

b) **Analysis of the content with regard to positive, neutral and negative utterances.** Those results will be presented on a **macro level**, e.g. displaying the overall results for a certain destination. This includes main terms occurring, the overall evaluation of the destination, main categories the region is classified in, language distribution, statistical information on the entries like word and segment counts, etc. On a **micro level** each blog entry is shown in a separate analysis, giving details on the evaluation, main utterances (positive, negative), length of entries, etc.

c) **Analysis of blogger characteristics.** One may also want to try to find out if the user writing the entry is male or female, his/her age and other socio-demographic data available through the blog or the user description.

The results will be provided in report forms, which should help users analysing different aspects of their tourism business.

In general, the type of analysis we intend to provide is called “sentiment analysis” (see Hatzivassiloglou, McKeown, 1997; Turney, 2002; Turny, Littman, 2003; Esuli. Sebastiani, 2006, 2007). Sentiment analysis has been applied to several fields, e.g. in the area of automotive or cinema films. Most approaches are linguistic based, thus a combination of computer linguistic methods, e.g. based on machine translation approaches and statistical methods to evaluate the results of the linguistic analysis. Other approaches dealing also with social networks and the analysis of the communication can be found in Stein et al. (2007) with a model for topic identification, Berendt & Draheim (2007) dealing with an e-mail corpus from a German news service using LSA (latent semantic analysis) and Hosser et al. (2007) showing how to analyse the development of topic trends in a newsgroup for mobile phones.
Technologies to be taken into account are:

- **Text classification**: assigning the document to certain categories, e.g. complaint, entry about a restaurant;
- **Domain classification**: assigning the document to a certain domain like meal or accommodation, wellness, etc.;
- **Text summarisation**: retrieving the most important segments of the document;
- **Language classification**: automatically assigning the language associated with a document, this is especially important for the pre-processing phase as otherwise the language has to be determined by hand.

From a principle point of view the blogs analysis process can be divided into several steps:

a) **Preparing the necessary search data**: In a preparatory manual steps the user defines the relevant terms, URLs, etc. he/she wants to have checked. This can also include that the user specifies the list of blogs or forums he/she wants to have checked. In addition, term lists can be associated with valuation information (importance of terms).

b) **Running the search engine crawler**: In a second step, the term tables from the preparatory step are fed into a search engine crawler, which is targeted towards identifying all relevant blogs and forums. Basically, this will result in a more or less huge list of URLs.

c) **Evaluating the URL search results**: In a third optional step the user may check the URL list of step b) removing the ones he/she thinks are irrelevant or the ones he/she thinks are the relevant ones.

d) **Retrieving the content of the URL list**: In this step the crawler retrieves all relevant content from the URL list. At this point in time, the text of the blogs and forums is available and written into a document management system or database system. In this project, the OASIS XML XLIFF (2007) format is used to represent the format. The format is extended with several new information categories (attributes, elements) in order to model the blog domain specific requirements (e.g. utterance polarities).

e) **Linguistic analysis**: Now, a linguistic analysis is applied. First, the documents get segmented. The segments are enriched with grammatical information (NP, VP, etc., type of segment, like question, exclamation, etc.). The segments will also be grouped into paragraphs. In addition, all the terms and multi-word terms are retrieved from the documents building frequency tables based on words and multi-words and links to the corresponding segments and documents. The terms also get marked if they appear in any of the term lists supplied by the user. The mark-up is added to the XLIFF formatted document.

f) **Classification of documents**: Based on the term list the documents are classified with regard to different classifications schemes. This will be based on the identified terms and associates different probabilities with the document for each category. The classification is also added to the XLIFF documents meta characteristics. Other meta information added are author, URL, time, etc. The meta characteristics may be enhanced during the whole process if new information can be detected.
g) **Identification of positive, neutral and negative utterances:** Now the segments are analysed with regard to their power and polarity resulting in statements and valuated statements. This is explained in more detail in the following section.

h) **Production of summary reports:** Based on the utterances various aggregation levels are produced, which themselves form the basic for different reports (e.g. quality management reports, etc.)

The following figure gives a basic overview about the application architecture:

![Application Architecture Diagram](image_url)

**Figure 2: Basic application architecture**

**Identifying relevant utterances**

A key feature of the system is the identification of the utterances in the system. Utterances can be qualified as positive, neutral or negative representing the **polarity** (POL) of the term (or phrase). In addition, each utterance has an associated **power** (strength). The power (POW) of an utterance is a number between 0 and 1 and represents the weight of a term with regard to given context. The context is important as attributes (mainly adjectives) change their meaning – power and polarity. The meaning of the word “high” differs depending if it appears in the context of “mountain” or “charges”. In addition, the word meaning also depends on subjective measures. “High mountains” may be valuated positive by a climber while for others this may induce negative feelings and fear.

A key issue is also identifying multi-words and phrases. Words like “very”, “not” have no meaning in itself, they have to be seen in the context they appear in. Thus, we also need a model that allows modifying the basic polarity and weighting of a term like “dirty” when they appear in the context of another word. A multi-term like “very dirty” must get a higher power than “dirty” appearing alone. The problem here is that we want to stretch the value of the
original terms depending on the modifying term. While this is easy with regard to polarity as this just negates the original polarity power, modification is more subtle as we need to remain in a range between 0 and 1. Hence, we require a power modification function (PMF), which modifies the original power in order to stay within the limit 0…1. One simple solution could be to provide a mapping table associating a power value with the modified power value, e.g., “very” [(0.1->0.2) … (0.5 -> 0.7), (0.9-> 1.0)]. Values in between can be estimated based on the lower and upper boundaries.

Examples are given in the following table:

<table>
<thead>
<tr>
<th>Term</th>
<th>Type</th>
<th>Polarity (POL)</th>
<th>Power (POW/PMF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dirty</td>
<td>Adj</td>
<td>Neg</td>
<td>0.6</td>
</tr>
<tr>
<td>very</td>
<td>Adv / Particle</td>
<td>neutral</td>
<td></td>
</tr>
<tr>
<td>not</td>
<td>Particle</td>
<td>Neg</td>
<td>-1.0</td>
</tr>
<tr>
<td>heavy</td>
<td>Adj</td>
<td>neutral</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Adv</td>
<td>Neg</td>
<td>[(0.1-&gt;0.2) … (0.5 -&gt; 0.7), (0.9-&gt; 1.0)]</td>
</tr>
<tr>
<td>cockroaches</td>
<td>Noun</td>
<td>Neg</td>
<td>0.7</td>
</tr>
<tr>
<td>mountain</td>
<td>Noun</td>
<td>neutral</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 1: Term valuation examples

Obviously the numbers used depend on various circumstances:

a) Terms may appear in different modifying contexts, e.g. “heavy”. In the context of “The meal was heavy” heavy is used as a property of meal, while in “The route description was heavily wrong”, “heavy” used as an adverb modifies the power of “wrong”. Also the term “high” can have a positive or negative polarity depending on the context – “high mountains” vs. “high charges”. Therefore, we may need different table entries depending on the context. This is also the reason why a linguistic analysis will supply hints how to identify word groups. The whole situation becomes increasingly complex as the power computation gets more complicated if you think of utterances like “very dirty”, “not very dirty”, “not dirty” involving adjectives with a basic negative polarity compared to utterances like “very nice”, “not nice” and “not very nice”. “very” as a modifier here works in one case as a amplifier of the basic power (“very dirty”), while together with “not” (“not very dirty”) it weakens the power towards a more positive end. “not nice” on the other is more negative than “not very nice”. Thus, computational tables are required to handle these interactions.

b) The user: different users may qualify different terms differently. This applies both to the traveller and the business user. A backpacker may associate a lower power to terms like “dirty, cockroaches” than a luxury traveller.

c) Term power also depends on the context of the segment and paragraph and the number of occurrences in the document.

Computing the valuation of an utterance basically implies determining all the relevant POW and POL values associated with an utterance. As a segment may contain several more or less
related utterances we need to define an atomic utterance, which further can be combined to segment utterances and paragraph utterances by adding the utterances of lower levels. Alternatively, the values can be computed for each level too. A hierarchical approach will mainly differ through the weighting applied (normalizing due to the segment resp. paragraph numbers).

**Statement modelling**

Although applying a first straight-forward approach, it could be sufficient to detect only word and word groups consisting of adjectives and nouns (noun phrases, NPs). As shown above, there are clear restrictions of this approach as the “real meaning” of NPs can be easily modified by verbs etc. In the etBlogsAnalysis approach statements are modelled as expression consisting of a predicate (e.g. a verb) defining a relationship and a number of arguments (mainly NPs). The arguments themselves consist of a subject and one or more objects associated to the subject.

Thus, a statement looks formally like that:

\[ S(\text{predicate, subject, object...}) \]

where subject and object are both NPs consisting of a noun and optionally one or more adjectives.

Based on this, a valuated statement is defined as

\[ VS(S, \text{polarity, power}) \]

where polarity and power is defined on the principles given in the previous chapter.

Here is an example from the utterance:

“The shower is calcified.”

\[ S(\text{is, calcified, shower}) \]

\[ VS((\text{is, calcified, shower}), -1, 0.5) \]

**Document Classification and Valuation**

A first important step is to determine the domain and relevance of a document. An entry where the writer just talks about a family meeting somewhere might not be of high interest for the tourism community. Therefore, a selection must be applied, which removes those documents that have no or only low relevance. This can be done by determining the domain of the document. This classification can be done by searching the produced term list for specific key terms, which are associated with domains and sub domains plus some relevance measure indication of the term for the given domain. Based on the terms detected and the relevancies aggregated, a sorted list of possible domains with their probabilities can be produced. Depending on a threshold, irrelevant documents are removed. An output of the classification could be that the entry is about hotels and that it deals with the meals served.

Once the valuated statements have been determined from a document, overall tables and statistics can be produced. Segments will normally form the smallest aggregation level. But even this case may cause problems if the author inserts several topics within one segment. The next level will be the paragraph level which will sum up all the VS from the segment.
level. Both segment and paragraph level can then form the basis of the document valuation. The valuations on these individual levels are easily computed by just summing up all the individual VSs. The final level will be a set of documents, which sum up the values of the individual documents.

The summary process is obviously influenced by the user’s goal. Assume a hotel manager responsible for the quality management who might be interested in mainly negative utterances in order to find out weaknesses. As he is not so much interested in the positive statements he will define his term list profile in such a way that he associates only very low positive power values for positive utterances while he will increase the power of the negative utterances. Thus, he will immediately find out where actions will have to be taken. In addition, he also will be interested in the individual negative utterances in order to find out more. Hence, he needs direct access to the individual blog entries and utterances.

The client manager on the other side will also be quite interested in the negative utterances while this might give him the chance either to write a correction if the utterance was wrong or immediately get into contact with the writer to thank him for his valuable hints and explain the actions based on the writer’s entry.

On the contrary, the marketing officer will not be so much interested in negative utterances. He will be more in favour of positive utterances – testimonials – which he could use on the website.

The product developer will be more interested in new combination of services (expressed through terms). Thus, he will have a look at less frequent but interesting new combinations of words.

The following figure shows some output as intended to be produced by the project:

---

![Hotel Evaluation](image)

**Figure 3:** Simple example of an individual analysis of a hotel mentioned in different blog entries.
4 Sterne? Nie und nimmer!

Wir haben 3 Tage in diesem Hotel verbracht. Der Service war unter anderem schlecht. Es war etwa 3 mal pro Tag, das Personal am Abendessen nachgefragt hatte. Die Zimmer waren auch eher klein und einfach. Immerhin war der Preis aber günstiger als in anderen Hotels in der Nähe.

Reference


Further steps

As described above, the application consists of various blocks, which are currently in implementation. It is planned to have a first version of the software available at the end of September 2007. For the moment, the languages to be supported are English and German. This will be followed by an evaluation phase where we apply the software to a huge number of blogs. This phase will also help to improve the software. Finally, a product version is planned for the end of February 2008.

References


