On the Status of the Automated Categorization of the Bavaria Portal's Internet Search Engine According to the Life Event Principle

Erich Weihs\textsuperscript{1} and Thomas Reuter\textsuperscript{2}

1. Introduction

Within the space of a few years, the internet has become an indispensable part of our society's infrastructure. The most important services provided by the internet are those of the search engines, which serve as universal interfaces to the digital world and supply users with the desired information in the form of a link list. Sociology professor Dr. Rudi Schmiede at the Technical University of Darmstadt maintains that we have long been dependent on the services of the search engines and thus are "definitely on our way to a Google society". However, technological design as well as the development of information and media competence will determine whether this term is going to stand for social progress or for the disabling of human judgment.

Although we are still at the beginning of the road from the search engine to the practical use of a "find engine" that also answers questions, nonetheless this indicates a way for search engines to help the computer-literate even more effectively thanks to new technologies. The search engine of the Bavaria portal (www.bayern.de) uses new technological methods to develop and process the public offerings (of government, authorities, municipalities, etc.) according to the life events of citizens.

The life event principle involves adjusting the online services of public administration, municipalities and other official bodies to the needs of citizens arising from concrete circumstances in their lives. In the process, all the information available from authorities and municipalities is intended to be offered to citizens in relation to a specific life event (moving, marriage, birth, etc.). This considerably enhances the service orientation and quality of the services. The life event principle is based on an agreement between the national and provincial governments which permits expansion or refinement. This is necessary for a state covering such a large area as Bavaria and has been implemented through such categories as "my environment" and "from my community", for instance.

Although we are not far enough along yet to answer questions thus formulated "intelligently" in practical operations, at least we can fan out the results list of a search according to the situation-related life events. The roughly 4½ million pages of Bavaria's public offerings are categorized as required for this purpose by the methods described in the following.

2. Google versus specialized search engines

With its market share of more than 90 percent, Google has practically achieved a monopoly. Today, "Google" is practically synonymous with the search machine per se thanks to its dominant position on the market and clever marketing strategies. There is no doubt that Google has an impressively large index (it

\textsuperscript{1} Erich Weihs, Insterburger Strasse 7, 81929 Munich, Germany, erich.weihs@web.de
\textsuperscript{2} Thomas Reuter, Hospelstr. 41, 35037 Marburg, Germany, th.reuter@cogisum.com
is alleged to have ten billion pages), but dissatisfaction with the search results is growing. Criticism is aimed primarily at the following points:

- unclear, often very high numbers of hits which appear plausible but cannot, in fact, be checked, since leafing through thousands of links is practically impossible
- watered-down hit quantities (spam, paid links)
- lack of transparency in the way the hits are ranked
- necessity of previous knowledge on what is being sought in order to be able to formulate search queries precisely by means of operators and further search terms.

Google itself confronts this criticism by offering special searches which only address specific segments of the internet in order to focus the hits. The special searches cover, for instance:

- news search
- image search
- space-related search for/in maps
- product search
- search in groups/forums
- search limited to scientific or scholarly sites
- search limited to local sites.

Segmenting the search space and search objects is a strategy for coping with the flood of information on the internet and is quite successful in certain areas, such as map search. Here the offerings of official agencies are of secondary importance, at least for citizen-centric uses.

The search engine of the Bavaria portal (SUMA) tackles the problem by starting with the segmentation of content and limiting the search space to the public sector (municipalities, educational institutions, employment offices, state ministries with subordinate agencies such as offices). In contrast to Google, the segmentation follows the life event principle described below.

3. The life event principle – environment-related aspects of life circumstances

3.1 Overview

Public authorities must satisfy the public's demands for information without structuring their response according to the perspective of the government agencies' own organizational structure. In the following, we will focus on environmental information in accordance with the Environmental Information Act (UIG):1

The Environmental Information Act, which took effect in 2005, obligates official agencies to provide environmental information as needed according to the life situation of citizens. This is all the more imperative considering that "environment" is not defined in the Environmental Information Act according to the jurisdictions of the federal or state administrations. What is more, the now ratified Aarhus Convention also entails this obligation.

The term "life events" refers to a sociological concept, cited in the introduction, in which the interaction of various economic, social and cultural, as well as environment-related, factors in the concrete life situations of individuals and social groups are theoretically comprehended. Apart from the objective, material and immaterial dimensions of a life event, the subjective dimensions are also addressed by relating only activities and decisions, but also interests and expectations of men and women to objective dimensions of their life circumstances. The concept has been introduced into policy analysis as "gender mainstrea-

---

1 Further laws, such as the Consumer Information Act, etc. must also be considered as EU directives, a "responsibility finder" or the services directive.
ming". The life event approach is a useful instrument of analysis for comprehending policy in relation to target groups and can thus afford a foundation for designing policies according to gender mainstreaming. Gender mainstreaming is already being variously implemented for classifying the internet services and information provided by public authorities.

Bund.de uses the life event principle to orient itself toward the results compiled together with Deutschland-Online. Deutschland-Online is the joint e-government strategy of the federal, state and municipal governments. It promotes an integrated e-government offer for all levels of administration through joint infrastructure and standards, as well as by networking all portals. Thus there is hope that this kind of classification will assert itself in time as an "interface" – the necessary condition for a search across portals.

3.2 Implementation in the Bavaria portal

The life events of the federal government are divided up for this purpose into the areas of "life event model for citizens" and "life events for companies". Later, there will also be life events for the administration. The definition of life events was set down in a federal-state work group. This definition leaves room for expansion. The organizing principle was taken over for the Bavarian search engine (SUMA). In order to include the municipalities adequately according to the Bavarian model, the life event principle must be expanded, for instance, by adding such categories as "life in my municipality", "tourism", etc. By the same token, "my environment" was added as a main category.

In the context of the considerations under discussion here, the life event concept is intended to enable the public to access the internet offers of public authorities according to the public's own needs. We assume that this organization is better tailored to needs than a presentation based on organizational charts. The life event principle will continue to develop according to objective criteria and facilitates the search in a manner "tailored to situations", while organizational principles, by contrast, can change abruptly at any time and are therefore less suitable for a search aimed at needs. The organization of the life event principle covers a large number of "environmental events", such as, for example:

<table>
<thead>
<tr>
<th>Construction</th>
<th>Animal welfare</th>
<th>Refuges, sanctuaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preservation of historic sites</td>
<td>Control of animal diseases</td>
<td>Lakes / bodies of water</td>
</tr>
<tr>
<td><strong>Society and politics</strong></td>
<td>Livestock breeding/animal husbandry</td>
<td>Road construction / road maintenance / traffic</td>
</tr>
<tr>
<td>Agenda21</td>
<td>Veterinary inspection</td>
<td>Environmental laws and regulations</td>
</tr>
<tr>
<td><strong>Danger prevention and security</strong></td>
<td><strong>Habitation</strong></td>
<td>Environmental information in the administrative district</td>
</tr>
<tr>
<td>Severe weather warnings</td>
<td>Supply and disposal</td>
<td><strong>Tourism in my municipality</strong></td>
</tr>
<tr>
<td>Civil defense and disaster prevention</td>
<td><strong>My environment</strong></td>
<td>Excursion destinations</td>
</tr>
<tr>
<td><strong>Health and preventative measures</strong></td>
<td>General</td>
<td>Recreation</td>
</tr>
<tr>
<td>Nutrition</td>
<td>Landscape conservation</td>
<td><strong>Agriculture and forestry</strong></td>
</tr>
<tr>
<td>Research and medicine</td>
<td>Special features of nature and the countryside</td>
<td>Information services</td>
</tr>
<tr>
<td>Consumer protection</td>
<td>Energy</td>
<td><strong>Forms</strong></td>
</tr>
<tr>
<td><strong>Animal husbandry and hunting</strong></td>
<td>Gardens / regional fruit products</td>
<td><strong>Research</strong></td>
</tr>
<tr>
<td>Protection of species</td>
<td>Groundwater</td>
<td></td>
</tr>
<tr>
<td>Exports and imports of animals</td>
<td>Maps, images</td>
<td></td>
</tr>
<tr>
<td>Fishing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 Gender mainstreaming means regularly giving consideration right from the start to the different life situations and interests of women and men in all social plans, since there is no gender-neutral reality. (http://www.gender-mainstreaming.net/)
3 "my environment" is lacking entirely at Bund.de (as of 1/2008).
Only environmental life events with highlighted main categories are shown. Corporate life events contain, for instance, the category of environmental impact assessment. As can be seen from Table 1, life events are structured hierarchically as a tree with three levels. The contents of sub-categories may overlap, such as Severe weather warnings, Measured data from ..., Information services. On the whole, about 650 categories must be administered.

4. The search space

The search space of the Bavarian search engine (SUMA) is limited by official offerings, municipalities, institutions of higher learning, employment offices, chambers of commerce, etc. In certain cases, additional institutions such as churches are included in order to take account of the regional supply of kindergartens, for instance. On the whole, approx. 5000 URLs are tracked, with approx. 4½ million indexed sites. The data material is extremely varied. Along with html pages, it also reads and completely indexes pdf and doc formats, among others. This means that such things as municipal bulletins and other reports consisting of several hundred pages are collected and completely indexed. While reports can sometimes be thematically circumscribed section by section, this is often not the case with bulletins such as community magazines, etc. Here it is frequently a question of a chronological or editorial mixture of news, reports and statements, whose content cannot be assigned in detail to the life event classifications.

5. The classification procedure (method)

In principle, there are two ways to divide up a search result list according to life events (or another classification):

- an ad hoc classification for the result list only
- a non-recurring classification by way of the indexed basic population of all links collected by the search engine before searches begin.

The ad hoc solution will not be pursued any further here, since the classification results depend on the search – and are not significant, statistically speaking, for short lists. Basically, the following procedures are possible:

- Marking the websites by the information provider, for example, by meta tags
- Defining (semantic) word profiles as a basis of categorization
- Mathematical-statistical methods
- Explicitly defining URLs to which a life event has already been assigned, for instance, when collected by the search engine
- Manually manipulating sites found by the search engine, for example, in the event of advertising, morally offensive material or in the ranking (not pursued any further here, reasonable in specific cases).
5.1 Meta tags

The most objectively certain classification for the operator is that undertaken by the data providers themselves. Classifying according to life event, space and time can be especially favorable for sites which must be newly created or for documentations of a web offering:

```html
<meta name="ByLebenslagen" content="ByLl_MeineUmwelt_Energie"/>
```

For the spatial reference system by place name (ZIP code, coordinates), the following syntax can be used:

```html
<meta name="ByRaumbezug" content="ByRb_GN[Amberg|Bogen]"/>
```

The same syntax can be used for the time reference. For pdf or MS-WORD documents, the meta tag is "hidden" in the document properties, in Word, for instance, under the category of "summary info".

5.2 Definition of word profiles

Defining word profiles by means of Boolean operators is one method for assigning the content to the appropriate categories of life events. Here simple semantic relations can be evaluated.

```xml
<Terms>
<QueryMatch>
<RelevanzString />
<Suchstring>((Database or catalog or information system) near[50] (Environment)) or ((Database or catalog or information system) near[50] (Health))

... ... ... ...
</Suchstring>
</QueryMatch>
</Terms>
```

The "A near$nn$ B" operator shown above only selects objects for which the terms A and B are a maximum of $nn$ words apart. The value of $nn$ must be ascertained empirically, it depends on the "language" of the chosen category. Looking at Fig. 1, it is immediately clear that there is an optimum number of hits for $nn$, because the maximum number of hits increases as $nn$ increases, up to the maximum number of hits possible, although the desired semantic context between A and B is lost. Experience has shown that values of $nn$ between 20 and 90 are useful for approximating as nearly as possible the characteristically ideal optimum according to Fig. 1.

---

1 Linguistic usage distinguishes here between, for instance, citizen and corporate life events.
5.3 Mathematical-statistical methods

We expect mathematical-statistical methods of classification to make it easier for us to constantly keep updating the categories through an "intelligent" automated process. While word profiles are sharply delineated (owing to the Boolean operations in the document of the class, the term must exist / cannot, etc.), classification is done through distance measures. The object is assigned to the class to which it has the least distance. This makes the method less sensitive when indexing an object for the first time, since the distance is calculated from all n-dimensional spatial coordinates.

Here the search machine of the Bavaria portal uses a support vector machine (SVM) for automatic categorization. The basis is a set of training objects for each of which the life event category they belong to is known. The training objects can be obtained from the internet either by semantic word profiling or manual selection. Each object is represented by a vector in a vector space. The task of "leaning" by the SVM method consists in choosing suitable coordinates to determine the space and inserting multidimensional hyperplanes which function as separation planes for matching the other objects. This method makes its possible to determine how precisely the objects being classified must be adjusted in order to be accepted into a life event (class)\(^1\) \(^2\). The reliability of the classification depends on

- how representative the objective contents of the learning documents are, which assumes a precise definition of the life events (there is still much to be done here), and
- the number of words in the objects to be classified, obviously, as pointed out under 4. Thus objects with several hundred pages of objective content are present to the same extent as those which evidence little structure from a representational point of view. The broad margin necessary is then no longer adequate for accurate classification.

\(^1\) http://www.wikipedia.org/wiki/Support-Vector-Machine
As Figure 2 shows, the word profiles (blue and red, left illustration) separate the categories in a linear manner prescribed by the definition of the profile. The system of coordinates is determined by the explicit word profile. With the SVM method, a nonlinear hyperplane separates the categories. The SVM method itself uses prescribed learning objects (here html pages or documents) to determine the best system of coordinates for effecting the separation.

5.4 Further methods

Practical experience has shown that it is advisable or even necessary to accept or reject URLs explicitly for specific categories, for example, the well-known, unmistakable URL of an information service. Along with the objective categorization according to events, web offerings are also classified by place using the Official Municipality Characteristic Numbers, place names, coordinates or meta tags as described in 5.1. This also enables a search to take the place into consideration.

Fig. 3: Optimizing the near parameters for waste water and fees
6. Results

At the time this report was written, roughly 4½ million pages had been indexed. The information providers were the Bavarian ministries with their subordinate service agencies and offices, institutions of higher learning, chambers of commerce, various nonprofit organizations, religious groups, etc. Of these, roughly 30% were classified, which is quite a respectable figure when you consider that many pages are only navigation pages, have contents which are too general or not classifiable, or consist of private-sector advertising. The boundaries are hazy, especially for municipal content (for example, vacations on a farm). Overall, 650 life events were used for classification, divided up into citizen life events and corporate life events. Sometimes, the available definitions, such as semantic definitions, meta tags and classification, were used in combination to define the class. During classification, it became clear that, depending on life event, one or the other method, or a combination, was most suitable and how stable the categorizations are when indexing an object for the first time.

6.1 Semantic definition

Fig. 3 shows the optimization of the near range from 2 to 99 words between the terms "fees" and "waste water" in the semantic definition for the main "My Environment" classes. A review of the category "My Community" shows that here the probability of obtaining useful hits decreases starting at about fifty words distance within the scope of the near command mentioned above. For instance, the topic of waste water is defined frequently but not semantically in the Community News category of the Community category. Community News have very far-reaching texts which often cannot be semantically defined to any reasonable degree. Quite frequently, this range of variation also proved successful with other definitions.

6.2 Automatic classification

![Fig. 4: Comparison of the semantic definition method and classification using the "support vector machine" (SVM) method](image)

Fig. 4: Comparison of the semantic definition method and classification using the "support vector machine" (SVM) method
The quality of the SVM method was checked in several classes by running separate classifications using each method and comparing the results (Figure 4). The SVM method requires between 50 and 200 learning pages in order to create significant classes. Originally, a manual search on the internet was used to find the learning documents for the classes and once found, prescribed to the SVM "for learning". This procedure proved to be too time-consuming, however. In addition, the search is too subjective. Instead, it turned out that searching pages using semantic definition and defining a foundation for learning by taking a random sample of from 500 to several thousand pages was more to the purpose.

Figure 4 compares the categories of energy and climate. It can immediately be seen that the distribution of the results agree very closely. By varying the near operator "Climate near[s-99] Climate" ensures that their contents are comparable.

7. Summary

For Bavaria portal, www.bayern.de, a "find machine" is used which covers the data space of public administration and closely related institutions, such as chambers of commerce, universities, employment offices, etc. What makes the find machine special is that it divides up the results lists according to life event. The life events are based on an agreement of a working group of the ministers of the interior at the federal and state levels. The life events are based on the gender mainstreaming approach, which categorizes by requirements and not in relation to a specific organization or a specific field. The find machine of the Bavaria portal can categorize automatically using the support vector method. The requirements for successful categorization are to specify about 50 to 200 learning pages which are obtained by semantic definitions. The advantage of the SVM method is enhanced stability of the results when the data space changes over time. Since the approach evaluates the distributions of the word space, it does not depend on the construct of a semantic definition, but instead only on the probability of its appearing. Another method which is increasingly being employed is to use a meta tag related to life event and place which enables a direct classification.

References

UAG Lagen (2006): "Gemeinsamer Vorschlag der UAG Lagen für DOL ("Joint Suggestion of UAG Lagen for DOL") (Juli 2008)" of the DOL sub-working group "Lagen" ("Situations") of July 12, 2006 - 1


Emmy Noether Project: Mining Lexical-Semantic Knowledge from Dynamic and Linguistic Sources and Integration into Question Answering for Discourse-Based Knowledge Acquisition in eLearning (http://www.ukp.tu-darmstadt.de/projects/qael/)


