Semantic Tagset Generation and Enrichment by Measuring Co-Occurrences in Online Social Networks

Julia Waltl

MASTERARBEIT

eingereicht am
Fachhochschul-Masterstudiengang

Interactive Media

in Hagenberg

im September 2014

Declaration

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they have been indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere.

Hagenberg, September 29, 2014

Julia Waltl

Contents

Kurzfassung

Das Web 2.0 bietet eine Vielzahl neuer Möglichkeiten. Das Hinzufügen von semantischen Metadaten wird durch semantic web APIs unterstützt. Diese APIs helfen dabei, Inhalte zu kategorisieren und diese für Maschinen lesbar zu machen. Das Social Web hingegen ist durchzogen von unsteten, schwer analysierbaren Inhalten.

Durch sinnvolle Nutzung der Tags und Hashtags, welche in sozialen online Medien verwendet werden, können auch Inhalte im Social Web kategorisiert werden. Das Empfehlen von Tags und Hashtags wird dadurch durch eine soziale Komponente gestützt. Diese Arbeit zeigt, wie eine solche Empfehlung von Tags durch eine Kombination von semantischen und sozialen Komponenten verbessert werden kann. Dies hat zum Ziel, dass Tags sowohl von Maschinen als auch von Menschen verstanden werden. Durch die Nutzung von Tags und Hashtags, welche schon in sozialen Netzwerken verwendet werden, kann eine höhere Reichweite von Inhalten erzielt werden.

Abstract

Web 2.0 offers a lot of possibilities. Adding semantic value to the Web 2.0 has become in increasingly harder challenge. Semantic web APIs help enriching the web with metadata in order to categorise and index content. The social web, on the other hand, is filled with noisy content which is hard to index. Through tagging and hashtagging this gap can be filled. Users annotate their posts on social media with tags and hashtags. By exploiting these tags and hashtags, tag recommendation on content is enriched with a social value. This thesis lines out how tag and hashtag recommendation can be achieved by combing the semantic and the social web. Through this, content can be annotated both in a machine-readably, and in a human-readable way. Using tags and hashtags which are already in use by an online community can increase the reach of posts.

Chapter 1

Introduction

1.1 Motivation

Through tagging, content can be categorised and indexed. Furthermore, relations to other content can be established. Information is organised efficiently and browsing through information becomes easier. This concept is widely used in online media and also social online media. Furthermore, authors assign tags to content such as articles before publishing them. Through efficient tagging content can perform better in terms of online search and exposure in social networks. Well chosen tags ensure that content is found by consumers who are interested in this topic. Additionally, authors use hashtags in social networks when content is published. Hashtags help collecting reactions and opinions from content consumers as well as they help distributing the content throughout social media.

In order to benefit from the properties of tagging, users of online social networks use tags and hashtags on a daily basis for their own content (e.g. text, audio, video or articles), or the content of others. As the target group already uses tags and hashtags to classify content, the content creator can take advantage of this. By paying attention to the tags and hashtags used by the audience of interest, the content creator could chose a better tag set for their content and ensure therefore, that the content is exposed and spread through the community.

As the social web moves rapidly, information spreads from user to user very fast as well. On the other hand, information can also get lost easily in the social web.

Information gets lost when it does not reach users who are opinion leaders. Recommendation by various opinion leaders leads to broader spreading of information.

Through determining the tag and hashtag language the audience of interest is using, the content creator can adjust the used tags and hashtags for their articles according to terms which are already used by the com-

munity. Due to information flooding, as is the case with the spread of the internet, information management has become so difficult that individually relevant information is drowned out by noise. To help filter relevant information from that excessive offer of information, tagging has emerged as a light weight, easy to use and efficient way of marking information with meta-information and hence creating semantics that yield to more relevant hits on web searches.

1.2 Objectives

This thesis aims to show how semantically generated tags can be enriched by using the collaborative tagging and hashtagging by a community throughout different social networks. In order to examine the validity of this approach, a tool has been developed which takes a generated tag set and enhances this tag set by scanning Twitter, Soundcloud and Mixcloud for co-occurring tags. This tool aims to assist content creators in the field of music journalism, therefore services as Soundcloud and Mixcloud are taken into account when the tag set is enhanced.

Soundcloud and Mixcloud are fast moving networks in the field of music publishing. In contrast Twitter is a micro-blogging service which uses hashtags. The thesis strives to determine the possibilities of the unification of tags used in music-centered networks which assemble folksonomies and hashtags used in a micro-blogging system.

Semantically generated tag sets are common practice to annotate tags to information. The thesis points out, how this approach works, which technologies are used by different semantic web APIs. The flaws and benefits of tag sets generated by semantic web APIs are examined.

Furthermore, it aims to line out different usage of tags and hashtags and how the usage influences recommendation of tags. Tag recommendation in social networks is mostly based on folksonomies in state-of-the-art research. This thesis discusses the gap between tag recommendation through folksonimies, and hashtag recommendation in social networks. Tagging in folksonomies and hashtagging in social networks differ from each other. This thesis aims to show approaches of unification of those two practices in order to generate benefits for content creators who distribute their content mainly through social networks.

1.3 Outline

First, in Chapter 2 semantic web technologies are introduced. This Chapter focuses on examining semantic APIs for automated annotation and concludes with an evaluation of those APIs. Chapter 3 deals with tagging behaviours throughout popular social networks. There, different social net-

working services and their usage are reviewed. Furthermore, tagging and/or hashtagging habits in those networks are examined. Thereafter, state-of-the-art practices in tag recommendation are discussed in Chapter 4. Those include various fields such as folksonomy-based recommendation systems as well as hashtag recommendation. An emphasis is put on the strategy of measuring co-occurences of tags, as this concept is used in the tag recommendation tool which is presented in Chapter 5. Chapter 6 concludes and evaluates previous, discussed contents, and gives an outlook on future work in this field.

Chapter 2

Semantic Web

The World Wide Web provides a large and constantly growing amount of information. This information is not only accessible everywhere, it is also easily accessible all over the world. Also the growing amount of different devices supports the spread of information. Approximately one in every five person owns a smartphone. With new mobile technologies at hand for a fifth of population, not only the number of content consumers increases, but also the number of content creators. The Web is moving rapidly, and new information and content becomes available with every second that passes.

Information is mainly prepared for humans readers, as the content creators are human as well. This proves to be a burden when it comes to classifying information and content for machines. Humans can easily comprehend, classify and arrange information which is completely incomprehensible for machines. Different languages, encodings and semantic technologies on websites add up to the problem of accumulating different information for one topic and processing this content.

In order to tackle those problems, the idea of a semantic web has been introduces by Berners-Lee, the director of W3C[1]. The proposed idea says, that data on websites shall be put in order through a structured and unified environment. He sees the semantic web as an extension of the existing web through which meaning is added to information. This shall lead to better cooperation of machines and humans [3].

2.1 Semantic Web Technologies

In order for machines to understand information, technologies have been introduced. Those are represented by the Semantic Web Stack (see Figure 2.1) which was also introduced by Berners-Lee in 2001. Higher layers are dependent on the layers underlying them. Major semantic web technologies are described in this Section.

[1] http://www.w3.org/

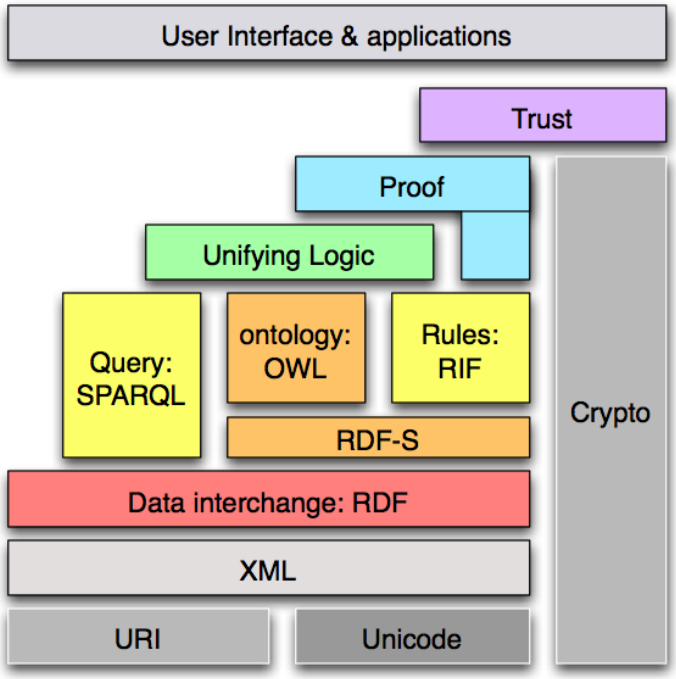


Figure 2.1: Semantic Web Stack [3].

2.1.1 URI

URI stands for Uniform Resource Identifier. A URI aims to provide a simple way of providing information about a resource. It is a unique sequence of characters, which explicitly describes a resource. This resource can be abstract or physical. Resources described by URI are not limited to web resources. They can also be cities, persons or companies. Familiar examples of resources which are identified by URI are electronic documents, images or a source of information with consistent purpose (e.g. the weather report for a specific city).

The syntax of URI is a sequence of hierarchical components. Those are referred to as the scheme, authority, path, query, and fragment. Though only scheme and path are compulsory. Figure 2.2 shows the syntax of URI [2].

2.1.2 XML

Extensible Markup Language (XML) is a markup language defined by W3C, which is human- and machine readable alike. XML aims to enhance simplicity, generality and usability throughout the internet. XML was initially designed for documents, but is also widely used in other fields such describing data structures or representing complex data. It is often used in web

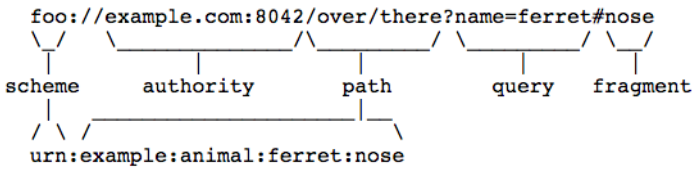


Figure 2.2: URI structure

services.

With the help of XML, metadata is added to a document. The metadata describes the resource by setting the documents' logical hierarchy and adding information to parts of the document. This is done via XML-Tags, for example:

```
<city>Linz, Upper Austria</city>
```

These tags can be defined for each requirement. Because of that, XML is a very flexible format. A lot of other markup-languages, such as XHTML, are based on the XML specifications. On the other hand, here machines can not comprehend the meaning of the tag "city". Neither can they find the connection between "city" and "state" by themselves. Nevertheless, XML provides a solid base for semantic web technologies such as RDF and OWL.

Due to the possibility of arbitrary tags in XML, conflicts can occur when merging XML documents as the same tags are used in both documents. In order to prevent those conflicts, namespaces can be introduced. A namespace is a Unified Resource Identifier. The URI is added to the element through an attribute which is called "xmlns". Optionally there can be a prefix which can be used for shortening URIs [9].

```
<element-name xmlns[:prefix]=URI> ... </element-name>
```

2.1.3 RDF

Through the Resource Description Framework (RDF) data and information can be structured and also enhanced by metadata. Furthermore, it provides the possibility of describing relations between information resources. It was designed by W3C with the purpose of representing information in the web.

An expression in RDF consists of triples, which include an object, a subject and a predicate. Those triplets are called RDF Graph. Subject and object can be seen as nodes, and the predicate can be seen as an directed and labelled arc, which connects object and subject. This is illustrated in Figure 2.3 [36, 4].

Subjects and objects can be represented by URI, but they can also be blank. Then the nodes can be thought of as variables. Predicates have to be represented by an URI string.

Figure 2.3: Representation of RDF Graph [4]

Due to the simplicity of the syntax RDF, it can be extended easily. RDF graphs can be built to express, form and transform data structures or knowledge representations [36].

RDF can be expressed with the help of XML, as seen in the code below.. Here, an abstract representation of RDF in XML format is given [42].

```
<rdf:Description rdf:about="subject">
<predicate rdf:resource="object" />
<predicate>literal value</predicate>
<rdf:Description>
```

RDF Schema

RDF comes with a vocabulary description language called RDF Schema. It is used for making statements about attributes such as properties, classes and containers. RDF Schema provides basic classes and properties, and defines how they shall be used. The Schema is an extension of RDF, and provides the possibility of building taxonomies and ontologies with the help of RDF [36].

2.1.4 OWL

Web Ontology Language (OWL) is an extention of RDF. This knowledge representation language is used for modelling complex ontologies and knowledge bases. OWL is highly expressive language. For example, it provides the possibility of creating new classes by combining existing classes. Furthermore, complex relations can be modelled. Figure 2.4 shows an example of transitive properties. A property is transitive if A is related to B through property P and B is related to C through property P, hence A is also related to C through property P. Figure 2.4 shows this through an example of ancestors [41].

Figure 2.4: Transitive Relations [41]

2.1.5 Linked Data

The concept of Linked Data is to interlink enriched data with external sources. By publishing this data, machine-readable data is provided for a large set of applications. Though the definition of a dataset is unique, and additional information is included through back-linking. In order to achieve those goals, the following policies are introduced by Berners-Lee [3]:

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (e.g. RDF).
- Include links to other URIs so that they can discover more things.

The *Linking Open Data Cloud* is the most widely known use case for Linked Data. W3C introduced the Linking Open Data project in 2007. Data without licence restriction is converted to RDF according to Linked Data principles and thereafter published. Everyone can participate in this project by providing RDF data and interlinking this data with existing entities in the Open Data Cloud.

Applications

As the Linked Data Cloud grows software is developed which are aiming to exploit the numerous possibilities of interlinked data [5]:

Linked Data Browsers: Those browsers enable the user to navigate through the complex depths of the Linked Data Cloud just as browsing through HTML content. In this case, the user progressively traverses the Web by following RDF rather than HTML links. The Tabulator browser[2] and the Marbles browser[3] are exampled for Linked Data Browsers. By tracking the origin of data and merging data about the same things, they display content from the Linked Data Cloud in a way which is

[2] http://www.w3.org/2005/ajar/tab
[3] http://mes.github.io/marbles/

comprehendible for humans. In Figure 2.5 the Marbles browser displays information on Tim Berners-Lee. Coloured dots indicate that data was merged.

Linked Data Search Engines: Human-oriented Linked Data search engines provide the user with relevant results according to the query. While having a similar interface as market-leading search engines such as Google[4], Linked Data search engines provide pursuing and detailed information on the underlying structure of found data.

Domain-specific Applications: Using selected datasets from the Linking Open Data Cloud, mashups can be generated. Those can be used for specific purposes only needed in special domains. The British Broadcasting Corporation[5] (BBC), for example, uses the Linked Data Cloud for matching topics throughout their data. This is necessary, as BBC uses different Content Management Systems for each of their stations. By linking the data to the Open Data Cloud, access throughout the network is given.

2.2 Semantic Web APIs

Semantic web APIs bring the web closer to the semantic approach. Various different APIs emerge, and though they reassemble one another, different ways of implementing semantic features are present.

Some APIs are basically text analyse tools, which take an unstructured text and put them into context by using technologies such as RDF, XML or JSON[6]. Through this, organising content becomes easier. Furthermore, the quality of information can be enhanced.

As it can be seen in Figure 2.6, semantic APIs enrich plain texts with metadata. The metadata used by the semantic API can vary, as classification systems differ. This will be outlined in he following, as three different semantic web APIs are introduced. The majority of semantic web APIs are commercial, though a free access is offered in limited form [11].

2.2.1 Alchemy API

The semantic web API Alchemy API[7] uses text mining for real-time text analysis. This API provides services such as sentiment analysis, keyword extraction, entity extraction, image tagging and more. Unstructured content is analysed in order to extract actionable information (see Figure 2.7[8]).

[4] https://www.google.com/
[5] http://www.bbc.com/
[6] JavaScript Object Notation
[7] http://www.alchemyapi.com/
[8] http://www.alchemyapi.com/

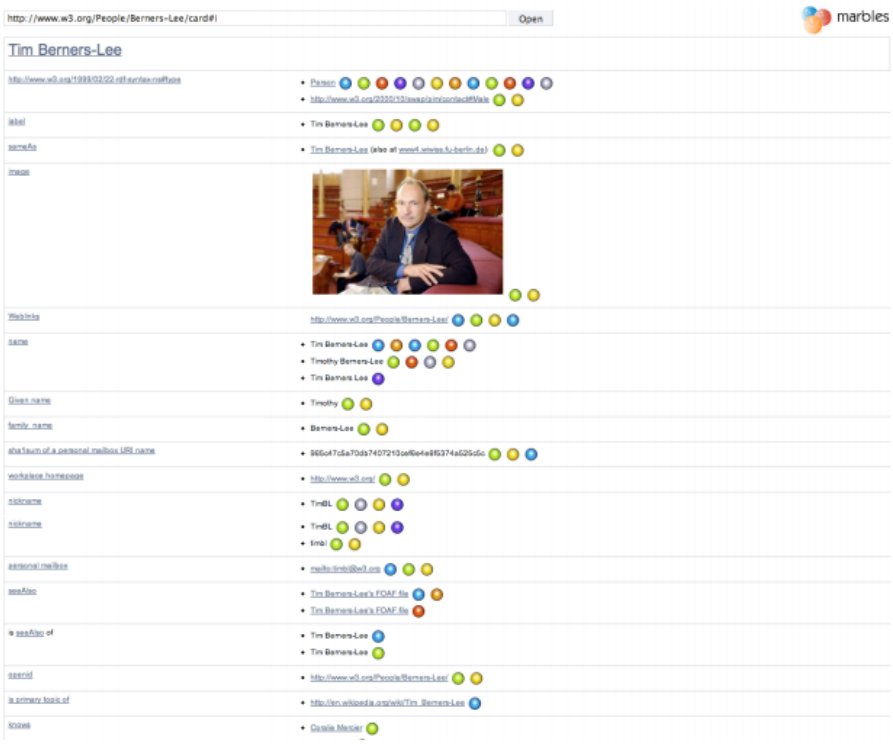


Figure 2.5: The Marbles Linked Data browser displaying data about Tim Berners-Lee. [5]

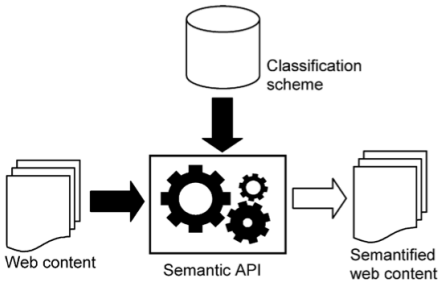


Figure 2.6: Semantic Web APIs workflow [11]

Counting an average of 65-75 million requests per day, Alchemy API is one of the leading semantic APIs. The majority of its costumers are social media monitoring firms, and 95% of the customers are paying customers [38].

Alchemy offers REST API endpoints as well as response meta-data in a variety of formats (XML, JSON, RDF). Also other languages than English are supported. Those include German, French, Italian, Portuguese, Spanish,

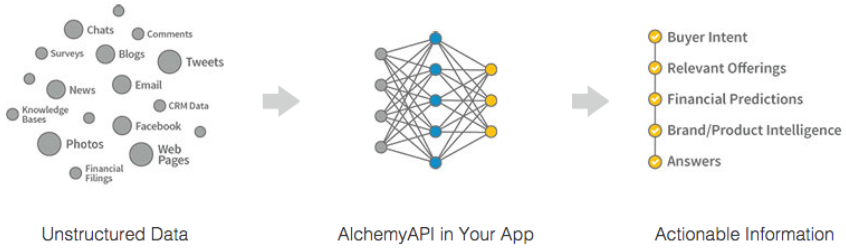


Figure 2.7: Workflow of Alchemy API

Swedish and Russian. When it comes to text analysis. following functionalities are supported:

Entity Extraction: The identification of people, companies, organizations, cities, geographic features and other typed entities is supported by Alchemy. This makes it easy to quickly set content into context and get hold of the main subject.

Sentiment Analysis: Sentiment is the feeling, opinion and emotion in a text. Alchemy provides multiple levels of sentiment, including document-level, entity-level, quotation-level, directional and keyword-level sentiment. A difference between positive, negative and neutral is made and a numerical value is added.

Keyword Extraction: When extraction keywords, the main topics and terms of a text are recognised. Those keywords can be used as tags for a content. Alchemy also ranks the extracted keywords. Those keywords are not limited to one word only, they can also be phrases.

Concept Tagging: This feature aims to understand the text as a human would. This is done by understanding how concepts relate. Through this, concepts which are not necessarily in the text can be extracted.

Relation Extraction: Here, subject, object and predicate are recognised within a sentence and their relationship to one another is looked at.

Taxonomy Classification: The text or HTML submitted will be classified into a hierarchical taxonomy, which can be up to 5 levels deep. Those levels represent the most likely topics categories.

Language detection: Alchemy supports the detection and classification of over 97 languages.

Text extraction: Plain text is extracted from websites. The text is free from any tags, advertisement or other unrelated content. The submitted text is cleaned.

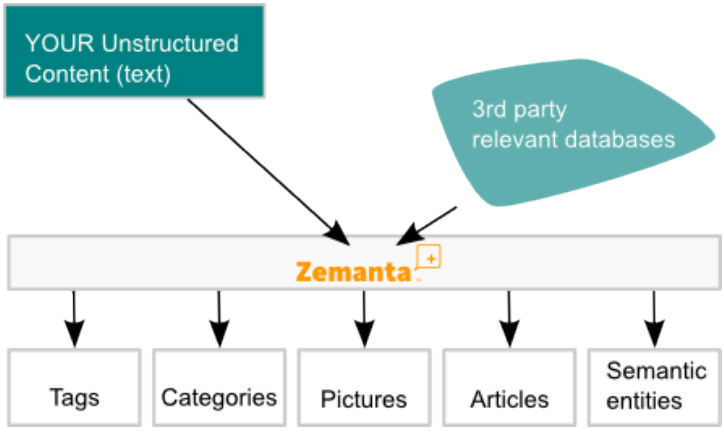


Figure 2.8: Features of Zemanta

2.2.2 Zemanta

Using natural language processing and semantic search, Zemanta[9] analyses user-generated contextually relevant content. Taking unstructured text, Zemanta extracts associated links, articles and images from the web. Furthermore, keywords, categories and tags are returned by the API (Figure 2.8[10]).

Through machine learning and data from other Zemanta users, the system can constantly improve. Zemanta API is at the time this thesis was written free of charge for up to 10,000 calls per day [11].

Articles: Associated content suggested by Zemanta is aggregated from news pages as well as blogs. The associated articles are presented as links.

Links: Looking at phrases and names from the texts, associated links are suggested. Most of the time those links refer to individual names such as persons or companies. The links refer to objects of popular databases, which include Wikipedia[11], YouTube[12], IMDB[13], Amazon.com[14] and more.

Keywords: Zemanta provides a finite amount of eight keywords for a text. Those are based on phrases and words from the text and also on related topics and concepts.

Categories: Categories can be longer phrases as well as single words. Cat-

[9] http://www.zemanta.com/
[10] http://code.zemanta.com/bostjan/clipart/api/general.png
[11] http://www.wikipedia.org/
[12] https://www.youtube.com/
[13] http://www.imdb.com/
[14] http://www.amazon.com/

egories aim to represent the main topics of the text. A text can be annotated with several categories.

2.2.3 OpenCalais

Open Calais[15] automatically generates semantic metadata for texts. This is done by using technologies such as machine learning and natural language processing. Extracted data includes entities, facts and events (see Figure 2.9[16]). Additionally, social tags are returned. The returned results are formatted in RDF.

Named Entities: As previously mentioned APIs, also Open Calais supports the extraction of named entities. This fundamental feature categorises the entities semantically (e.g. city, company, continent, person, etc.). Furthermore, URIs are returned for each found entity. Those contain a back link to the calais repository. From there, further information on the entity can be found in other databases such as DBpedia[17] or Reuters[18].

Facts and Events: Facts refer to positions, education statuses or similar descriptive facts. Those facts are extracted for events as well as entities. Events can inter alia refer to a release, an acquisition or a date.

Social Tags: This feature aims to tag the text as a person would tag it. The tags returned are mainly taken from Wikipedia entries. By finding generally valid tags, categorisation shall be made easier.

2.2.4 Applications

Semantic APIs provide a wide range of possibilities for making the web more semantic and comprehendible. Unstructured texts of all kinds can be enriched with important metadata through their usage. Many use cases take place in the section of content creation and content categorisation. Below some use cases are introduced.

Enrichment of Articles

For content creators in the web, a real-time analysis of their articles provides a lot of advantages. Related content, topics and concepts help them adding additional information to their articles. Enrichment of web articles are therefore a wide-spread use case for semantic APIs [1].

[15] http://www.opencalais.com/
[16] http://www.opencalais.com/about
[17] http://dbpedia.org/
[18] http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-index-and-definitions#acquisition

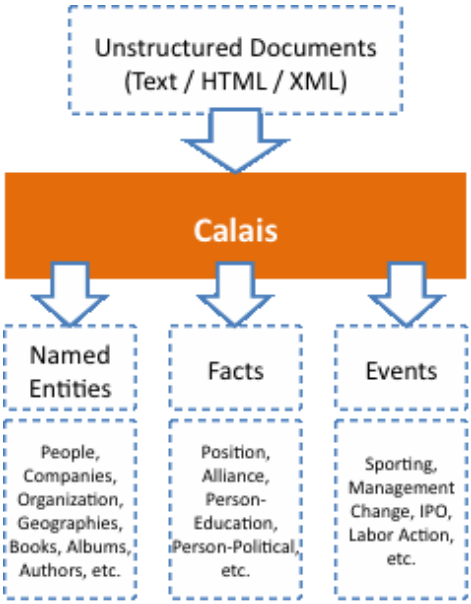


Figure 2.9: Extraction supported by Open Calais

By widgets provided by semantic APIs, images, links and tags can be added to content easily. In Content Management Systems (CMS), semantic widgets are popular amongst content creators. But also platform-independent browser extensions are provided by various APIs. Zemanta, for example, has developed both a plugin for the CMS WordPress[19] and browser plug-ins for the most commonly used browsers.

In Figure 2.10[20] the Zemanta Wordpress plugin is pictured. Here, related articles and images are suggested to the term *Ferrari* to the user. This takes place in real-time.

Semantic Search Engines

Through special applications such as Calais Marmoset by OpenCalais or AlchemySEO by Alchemy, texts on websites can be enhanced with meaningful metadata. This ensures, that the content is comprehendible for search engines. The search engines are provided with intelligent metadata, but the content seen by the users remains the same.

Mashups

Links provided by semantic APIs can be used to extend datasets (2.2.2). By accessing this data, mashups can be generated. Mashups combine multiple

[19] http://wordpress.org/
[20] https://wordpress.org/plugins/zemanta/

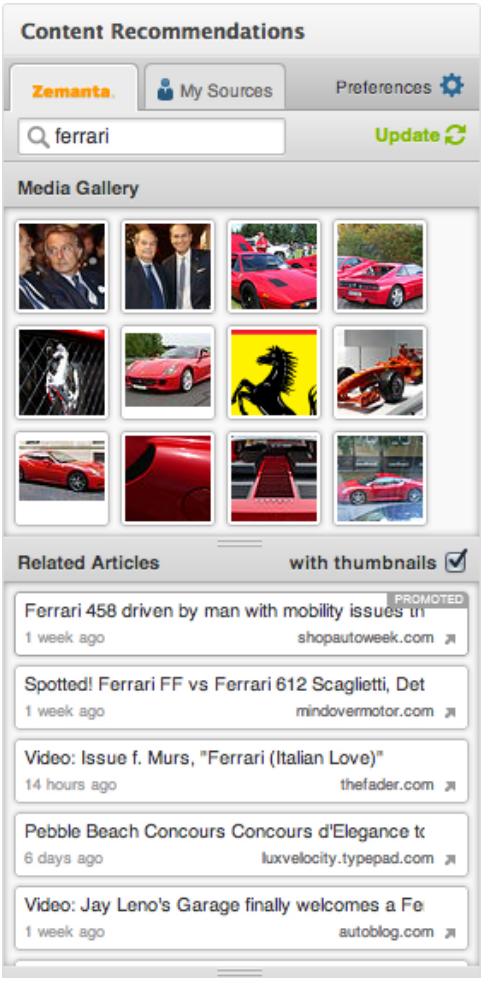


Figure 2.10: Zemanta WordPress Plugin

web resources and connect this data. Examples are Flickcurl[21], which gives RDF descriptions for photos on Flickr[22] or the RDF book mashup [6], which wraps up several book related APIs.

Analysis of Social Media Posts

Another broad field of application for semantic APIs is the analysis of posts on social networks. Most common use case is sentiment analysis. This means that through analysis the mood a post carries is extracted.
Users of social media often talk about their own moods and opinions. Posts carry both an expression of the authors' mood as well as their feelings to-

[21] http://librdf.org/flickcurl/
[22] https://www.flickr.com/

wards subjects. When talking about a subject, the mood is expressed more generally though [8].

As one study by Jansen ([14]) suggests, 19% of all posts on twitter, so called "tweets", are about or directed towards brands or consumer products. Out of those tweets, 20% contain a mood. This indicates that there is arguably a broad market for sentiment analysis of social media posts. By analysing large quantities of posts, an overall mood towards a brand or a topic can be generated. Often, semantic web APIs provide a sentiment analysis. Alchemy, for instance, offers a sentiment analysis API.

2.3 Evaluation of Semantic Web APIs

2.3.1 OpenCalais

Out of the APIs discussed in this Section, OpenCalais is an industry leader. It has been widely adopted by the open source community. The service is used by a lot of applications, and therefore likely to enhance and adapt fast according to the users' needs.

Extracted entities are generally of good quality, but entity disambiguation linking to open data datasets is lacking. This feature is provided for a small dataset only, which contain companies, geographies and electronic products. The extracted "Social Tags" and "Facts and Events" add additional value to the result set. Those results, though, are also not interlinked to a linked data service.

2.3.2 Zemanta

Zemanta is mainly marketed as a blog enhancement product, targeted towards bloggers. However, the semantic API provided by Zemanta is a good tool for extracting named entities and related content. This even can be done in a single API call. By calling `zemanta.suggest` high quality ambiguous keywords are returned. Furthermore, here several links to open databases are provided. Those include FreeBase[23] and DBpedia. One drawback of Zemanta API is, that the result set is limited to eight entries only. As those eight results are of high quality and relevant, a larger result set is not necessary in most cases.

2.3.3 Alchemy

Next to named entity extraction, the API provided by Alchemy contains a set of convenient features such as language detection, quotation extraction and content scraping and structured data extraction. Furthermore, this

[23] http://www.freebase.com/

API has no difficulties processing HTML and is even stripping unnecessary content (such as advertisement). Also, Alchemy processes scanned images.

When it comes to entity extraction, Alchemy convinces with faster response time than alternatives. From time to time, disambiguated URIs for named entities are missing. Apart from few exceptions, entities are always disambiguated. The number of results returned is rather low compared to other APIs [39].

Chapter 3

Social Network Services and Tagging

3.1 Characteristics of Social Networking Services

In order to examine different ways of annotating content in various Social Networking Services (SNS), it is important to understand their workings. In the following passage, a selected overview of classification of SNS is given [13]:

3.1.1 Community-Based SNS

Community-Based SNS are defined by text based interactions, which take place in virtual online communities. Those communities are based on common interests or objectives. This type of SNS can be classified as a type of content-based SNS. Community-based SNS provide the possibility of forming communities and sharing content with those communities. The shared content is user-generated and/or user-curated.

Due to those characteristics it can be derived, that platforms such as Facebook[1] and Google+[2] can be accounted into the group of community-based SNS. Both of mentioned platforms encourage the user to build communities.

3.1.2 Micro-Blogging

The most famous and wide-spread example for a Micro-Blogging website is Twitter[3]. This kind of content-based SNS is gaining popularity and users mainly but not only due to quick information distribution. Another reason for increasing usage of those kind of networks is its usage as a mean of

[1] http://www.facebook.com
[2] http://plus.google.com
[3] http://twitter.com

communication in political campaigns, activisim and disaster management. Interaction on Micro-Blogging platforms is high. In an online survey carried out by Bentcheva in 2013, 50% of the 587 participants stated to read their Twitter feed more than once a day. Furthermore, approximately 25% of the participants indicate to post more often than once a day [7].

3.1.3 Media-Based SNS

Here, a connection between users is created though various media formats. In comparison to content-based SNS, more interactions take place in media-based SNS. One kind of media-based SNS is photo/video/audio sharing. Another form of media-based SNS is virtual reality, see Figure 3.1. Prominent examples for Media-Based SNS are Flickr[4] and Instagram[5] for photo sharing, as well as YouTube[6] for video sharing. Considering proposed characteristics, Soundcloud[7] can be added to this list as an example of an audio sharing platform. In a limited way, also images can be shared on Soundcloud. Photos from the users' Instagram accounts can be assigned to a track posted on Soundcloud [26].

Further types of different types of SNS can be seen in Figure 3.1. This framework proposed lines out how different types of SNS are related to one another. It can be seen, that tagging is a core concept of Social Networking Services.

3.1.4 Social Media Streams

The previous Section has shown, that distinct kinds of Social Networks have different use cases. Therefore, user interaction patterns are different depending on which network they are using, and what purpose it is serving. The social media stream on the platform is shaped by the users' interactions.

Spectrum of Social Media Streams

Depending on the purpose of a Social Media Site, Social Media Streams can develop different characteristics. Some of them are lined out here [8]:

Interest-graph media: By following persons or organisations based on interests, an interest-graph is formed. Those interests can be mutual, for example following a news paper or person of interest is biased mostly. Furthermore, a connection in real life is mostly not required, as "follower" relationships are based on shared interests.

[4] http://www.flickr.com
[5] http://www.instagram.com
[6] http://www.youtube.com
[7] http://www.soundcloud.com

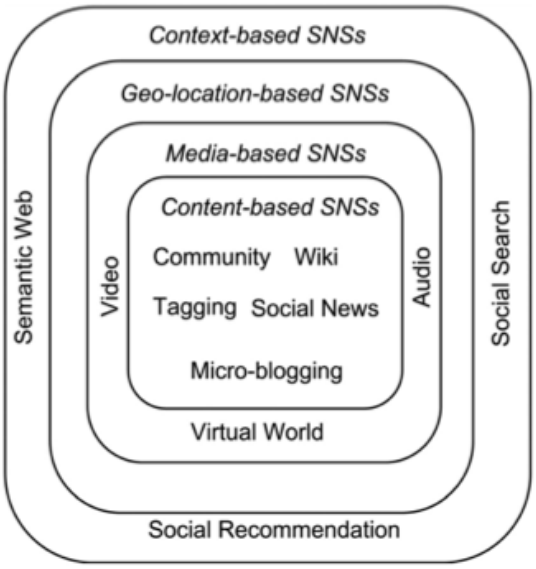


Figure 3.1: SNS framework [13].

Social Networking Sites are platforms which encourage users to form a reflection of their real-world social network. Basically, relationships which already exist in real life are transferred to a digital environment. A relationship of a connection which was established offline is extended and reflected in an virtual environment. Though connections are also made without, very rare or spare previous real life encounter.

Professional Networking Sites (PNS) provide a networking service in the context of work. Here, professionals can connect and recommend work contacts for others. Connections are interest based, and used for professional purposes. LinkedIn[8] or Xing[9] are popular examples for PNS.

Content Sharing and Discussion Services Those include video-, audio- and information-sharing platforms such as various Blogs or forums as well as YouTube/Vimeo[10] for video sharing or SlideShare[11] for information sharing. As the title implies, content is shared and discussed on those platforms. This content can be but is not necessarily user-generated or user-modified.

[8] https://www.linkedin.com/
[9] http://www.xing.com/
[10] http://www.vimeo.com
[11] http://www.slideshare.com

Social Media Stream Characteristics

Semantic interpretation of social media content can be difficult due to their broad and complex characteristics. State-of-the-art automatic semantic annotation is based on long, carefully written web content. Social media streams on the other hand are commonly short, inter-connected and can contain different kinds of slang-language. Due to challenging new characteristics, new technological approaches of exploiting social media streams have to be found. Main characteristics are introduced here [8]:

Short Messages or microtexts include very short messages posted inter alia on Facebook or Twitter. Twitter messages are restricted to 140 characters, and this restriction is exploited by the majority of tweets. On Facebook, there is a limit of 60 000 characters per post. But as a research by quintly[12], a social media analytics blog, shows, the Facebook post distribution per length has its peak at 2 characters per post (which the authors trace back to the usage of emoticons, which usually consist of 2 characters) and the second highest peak at 109 characters per post. Google+ officially has no limit on their posts. On this social networking site, post distribution by character has its peak at 156 characters [27].

Through this, the main traffic on social networking sites is made up by content which is shorter than average web content which is being used for semantic annotation.

Noisy content includes slang-like spelling, irregular capitalisation (all capital or all lowercase letters), emoticons and abbreviations. Emoticons are used as sentiment indicators. Capitalisation normalisation techniques have already been developed. Furthermore, various shortening styles in micro texts have been researched and techniques of normalisation have emerged out of research in this field [12]. Nevertheless, it is quite difficult for algorithms to detect differences in slang language styles based on the users' origin.

Temporal: User-generated content is exposed to a large audience in real-time, and also interactions take place within a short amount of time. So social media content is exhibited to great temporal dynamics. This means that popularity of content can grow and fade over time, in various intervals [24]. Therefore, it is important to take temporal patterns into account when examining content spread on social media sites.

Social context: Depending on the network, the content posted by the user is shared with a different network. On interest-based SNS, the audience differs compared to the audience in a community-based SNS. Therefore, the social context is important for interpreting content. Some factors have to be taken into account. Those include the network the

[12] https://www.quintly.com

user is posting the content on, who the content is shared with. The frequency of human interaction with content can give insights on how the content performs in the environment it was posted to. E.g., if the content is shared again by its recipients, it might be relevant to a larger group of users.

User-generated: As mentioned previously in this Section, users are producers and consumers of social media content. Either content which is generated by the user is shared (media, ideas, opinions) or the shared content is at least curated by the user (blog posts, news articles). Sometimes also a comment or opinion is added by the user who shares content from a third party. Demographic information about the user as well as interests and opinions can be mined out of user profiles. Mining user profiles can help to build up a shadow profile based on data which was shared by the user. Through that, preferences can be determined. By having knowledge about the users' preferences, user-generated content might be easier to classify.

Multilingual content creation. Less than 50% of all tweets are in English, but semantic technology methods have focused on the english language mostly [10]. So analysing all content exposed on social media sites proves to be difficult, as very few languages have been taken into account by semantic technology methods.

3.2 Tagging in Social Media

Tagging is a common way of indexing one's content in online social media. Not only do tags make it easier to search shared content, they also provide an efficient method of discover related content. Tags and hashtags are used throughout different online social networks [45].

3.2.1 Origins of Tagging

Tagging

Tags are used throughout information technologies in various areas, such as databases. In this context, tags are used for classification, marking ownership or noting boundaries. In social networks, tags first occurred in 2003 on the website Delicious[13]. On this bookmarking platform, users are provided the possibility of adding custom tags to their bookmarks. Also Flickr was amongst the first services to introduce tags. The concept popularised quickly, and other platforms such as YouTube or Last.fm implemented tagging.

[13] https://delicious.com/

Hashtagging

Hashtags have already been used in Internet Relay Chats[14] in 1988. Their purpose was, as it is today, to group conversations and content. This makes it easy to find associated content to a certain topic.

In contrast to tags, hashtags specifically start with the leading sign #. The first appearance of a hashtag on Twitter was in 2007. The usage of hashtags became a practice internationally during 2009 and 2010. In 2009 Twitter started to hyperlink all hashtags in tweets to their search results. Shortly afterwards, in 2010, "Trending Topics", a collection of hashtags which become popular within a short amount of time on Twitter, were displayed on their front page. Though Twitter may be the origin of hashtags in popular culture, hashtags are used throughout the world of online social media [28]. Services using hashtags include Google+, Facebook or Instagram. By indexing shared entities with hashtags, those entities will be exposed to a community, which shows interest in one or more of the used hashtags. Through linked hashtags, users can easily be part of an interest-based community.

3.2.2 Tagging Motivation and Strategies

General tagging motivations are tagging in order to organise personal collections and tagging for special purposes. These two kinds can also be labelled as categorisers and describers. While Categorizers are focused in organising information organisation-oriented and develop a personal structured tagging system, Describers are social-oriented. There are numerous describers on a single item. Through that, discovery and the amount of shares of an item is promoted. Additionally, tags in this content can serve as a form of expression by showing for example personal taste, preference or judgement [19].

Basically, an entity can be tagged based on its physical/objective attributes or based on the tagger's perception and judgement. So two major types of tagging strategies can be defined: object-based strategies (see Table 3.1), which are based on the characteristics of the entity, and situation-based strategies (see Table 3.2), which describe taggers' concept of the entity.

Object-based tagging is mainly used by Categorizers to describe the items' properties. Though Categorizers mostly use tagging as a personal strategy, so the tags are not necessarily fully objective. So also in a categorizers' scenario the taggers' concept plays a role. Describers tag content based on their perception. Through multiple Describers on a single entity, a folksonomy can be created.

[14] http://en.wikipedia.org/wiki/Internet_Relay_Chat

Table 3.1: Object-based Tagging Strategies [19]

Strategy	Example
By topic	“election2014”
By media format	“video”, “article”
By author or owner	“G. Martin” for articles about/from George Martin
By copyright	“open source”, “creative common”, “free”
By date/time	“july2014” for organising purposes

Table 3.2: Situation-based Tagging Strategies [19]

Strategy	Example
personal judgement	“fun”, “boring”, “cool”
self reference	“mystuff”, for organising purposes
personal task	“toread”, “travel”
symbols or numbers	hearts for liking number of rating
personal character strings, which do not make sense to anyone else	

3.3 Tagging and Hashtagging in Different Social Networks

3.3.1 Facebook

In June 2013 Facebook introduced linked Hashtags [29]. When clicking on a hashtag, Facebook redirects to a feed, which shows posts containing the hashtag. Furthermore, hashtags originating from other services such as Instagram are clickable too. Every hashtag has a unique URL.

Unfortunately, hashtags on Facebook can not be accessed by developers at the time this thesis was written [30]. Facebook provides the Graph Search API, which allows developers to search public posts. When submitting a hashtag query to this service, the hashtag sign in front of the term will be ignored.

However, the feature of adding hashtags on Facebook has been discussed widely, and is subject of controversial debate. One reason for controversial opinions on the feature is the lack of integration in the API as mentioned previously. Furthermore, critics state that the feature arrived too late, and user behaviour on Facebook does not necessarily need hashtags. Additionally, Facebook’s privacy structure makes hashtags less useful as they are on Twitter or Instagram. This is due to numerous possible sharing settings on every post. The social-media analytics blog quintly analysed over 38 million Facebook posts, and discovered that the usage of hashtags on Face-

book stagnates at 16% [46]. Furthermore, all-over engagement and interest in Facebook hashtags is rather low compared to other networks which provide a hashtag feature. Google Trends shows, that the interest in Facebook hashtags shows a peak around the release date in early 2013, but mostly the interest is lower than the interest in hashtags on other platforms [35].

The social media monitoring tool EdgeRank[15] discussed the impact of hashtags on exposure in pages' Facebook posts. They analysed their clients' data and found out that on Facebook the usage of hashtags does not lead to additional exposure of content compared to not using hashtags. They also found that viral reach is even higher when no hashtags are used [40].

3.3.2 Google+

Google+ already introduced hashtags in October 2011. An autocompletion feature was added in early 2012 [31]. On Google+, hashtags are automatically added to the users' content, provided the post has a sufficient amount of text. Those can be edited and/or deleted by the user. Those hashtags are built into Google's search, which makes search results for hashtags more precise. Furthermore, Google search provides a feature called "related hashtags" when searching for a hashtag.

Public Google+ posts can be searched by using the activity search API [32]. Co-occurring hashtags are not explicitly listed in the results returned by the activity search API.

3.3.3 Twitter

On Twitter, hashtags are an essential part of communication. As the hashtag as it is widely used nowadays gained its popularity on Twitter, this does not come as a surprise. Tweets containing hashtags get twice as much engagement compared to tweets which do not include any hashtags. Furthermore, tweets with one to two hashtags are more likely to be retweeted. Though the usage of more than two hashtags leads to a drop in engagement [43, 44].

Twitter's APIs make it easy for developers to search for hashtags and retrieve a list of co-occurring hashtags for each entity (tweet). Furthermore, Twitter provides the possibility of a streaming API. This means, that an HTTP connection to the Twitter servers remains open, and a live update of tweets containing specified hashtags can be retrieved [33].

3.3.4 Soundcloud

On Soundcloud, tracks are being tagged by the user who uploads the track. Soundcloud advises their users to use as many tags as possible, without polluting the tag list too much. Through the Soundcloud search API tracks

[15] http://edgerankchecker.com/

can be searched by tag. The returned tracks include a tag-list, which makes it easy to reveal co-occurring tags.

3.3.5 LastFM

At LastFM[16] tracks can be annotated with tags by every user. So a tag count for each track is available. In order to retrieve commonly used tags, the API provides a method which is called `Track.getTopTags`. As the method name indicates, the tags with the highest tag count are returned for the track that is submitted.

3.3.6 Mixcloud

Also on Mixcloud[17] user can add tags to the tracks they upload. The Mixcloud API also provides a search function which makes it easy to retrieve tracks which are tagged with a specific tag. Also, a list of all associated tags is returned by the service.

[16] http://www.last.fm
[17] http://www.mixcloud.com/

Chapter 4

Tag Recommendation through Social Networks

As stated in the previous Chapter, social networks have different characteristics according to their use cases. It has also been pointed out that the concept of tagging differs throughout social networks. Relationships between users differ, and so do ways of communication. This has to be taken into account when using tags from social networks for tag recommendation.

4.1 Tag Recommendation in Folksonomies

4.1.1 Folksonomy

The term folksonomy is a combination of the words "folks" and "taxonomy". This system of classification is a collaborative method of creating tags in order to categorise content. Classification of data is done by users in a folksonomy. Three basic items define a folksonomy: user, tag and resource. Each resource can by annotated by users with tags. There is a distinction between broad and narrow folksonomies [50].

Broad Folksonomy

A popular example for a broad folksonomy is the social bookmarking tool Delicious. In this context, numerous users are tagging the same object, and each user can use their own vocabulary.

By creating and uploading content, the user exposes the content to other users, which are able to tag the content as they wish. This is illustrated in Figure 4.2 (a). Here, both categorising and describing tagging motivations can occur. Users can see tags already annotated by other users. They use those tags as a resource for information to generate other tags to add. So, information for own tags is partly based on tags already annotated to the

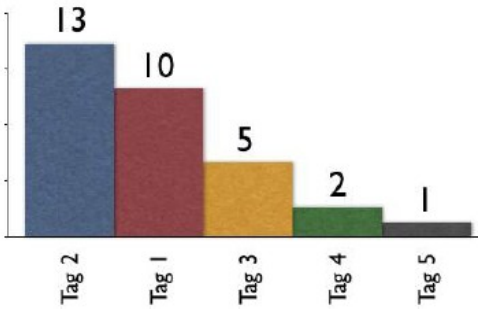


Figure 4.1: Low Power Curve forming a Long Tail [50].

content. Additionally, previous tags are used to find more resources associated with the annotated content.

Each user tags the object in different ways, depending on their own vocabulary, mental models and structuring processes. Also, the same tags often occur multiple times in a broad folksonomy. Other typical problems which occur in folksonomies are the usage of colloquial terms, spelling differences (e.g. "hard rock" and "hard-rock") or differences in language. Therefore, the number of tags for each resource in a folksonomy is rather high. Through this, broad folksonomies are tend to show a low power curve and a long tail effect:

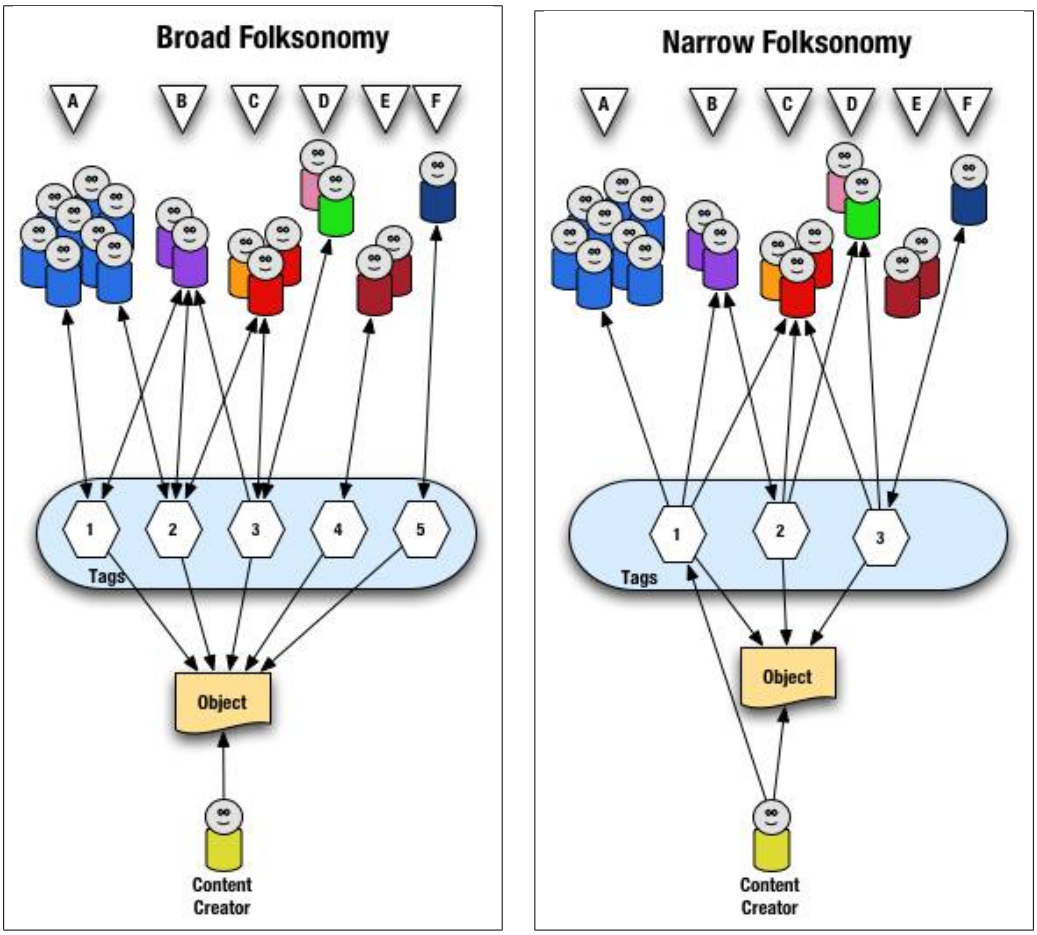
The low power curve is graphically illustrated in Figure 4.1. The curve reveals that a lot of users are using the same popular tags. But also smaller groups of users prefer more specified vocabulary for annotating the content according to their needs. This leads to a long tail[1] effect. As seen in Figure 4.1, tag 2 is annotated 13 times by various users, tag 5 on the other hand is only annotated by one user. Through those tag counts the long tail is created.

One can argue, that those effects are positive sides of broad folksonomies. First, there is the power of the quantity, so popular tags are promoted and trends in large groups can be investigated. Second, more specific search is possible through the long tail [48, 50].

Narrow Folksonomy

Flickr is one typical example of a narrow folksonomy. Here, an item is annotated by a smaller group of users and their motives are mainly their own convenience and future reference (categorising behaviour). Furthermore, information from previous tags is used for finding related content. The tags are singular in nature. This means, that one tag used by multiple users only counts as one tag. As Figure 4.2 (b) shows, the content creator often adds

[1] http://en.wikipedia.org/wiki/Long_tail



(a) & (b)

Figure 4.2: Broad & Narrow Folksonomy [50].

some annotations in the beginning, in order to get the tagging started. From there a small group of users adds some more tags.

Apparently the narrow folksonomy loses the richness of quantity, as annotation is carried out by a smaller number of users. Advantages are the usage of a more specific vocabulary, as the annotators are often part of the same interest group (i.e. professional portrait photographers on Flickr) which use a shared professional language. Through this, content search and retrieval is fast and efficient [50].

Properties of Folksonomies

As folksonomies provide numerous options to take advantage of socially generated knowledge, it is a widely discussed topic. In order to sum up the

main properties of folksonomies, advantages and drawbacks are listed here [48]:

Drawbacks

- Through different vocabulary and variability in language styles precision of tags is not always given. Furthermore, this leads to ambiguity: words have more than one meaning, or multiple words have the same meaning.
- Especially in a broad folksonomy chances of having too many tags to view them all are high. Also, numerous tags could lead to an information overload.
- In a typical folksonomy, the tags are not given a hierarchy. This makes it harder to index tags and group them accordingly. A Folksonomy is a flat space of keywords with no hierarchy.

Not all of the drawbacks are necessary a limitation to a folksonomy. Approaches such as normalising the tag set through providing pre-defined tags would change the nature of a folksonomy. And the features of a folksonomy propose a lot of advantages as well:

- As folksonomies are user generated, they reflect the opinion and conceptual model of the users annotation the item.
- Tags in folksonomies follow the users' language and terminology, therefore they match the users' capabilities and needs. Furthermore, this make searching and browsing information easier for users.
- Another property of folksonomies is their inclusiveness. There is no authority controlling added tags or filtering those.
- Through the long tail, discovery of information is enhanced. Browsing content based on rarer keywords can lead to incidental discovery of content. This concept is called serendipity.

4.1.2 Tag Recommendation in Folksonomies

Folksonomies contain three main properties (user, tag, resource). In Recommendation Systems usually a tag is recommended to a user for a certain resource. This approach most of the time is a *personalised recommendation* approach, which means, that the users' preferences are taken into account.

Another approach is to recommend tags *for* a resource. This approach is not targeted towards a specific user. Therefore, user preferences are not taken into account directly, and recommendation is mainly based on the resources' content and related resources. This can be seen as a *non-personalised* approach [22].

Personalised Recommendation

In folksonomies, a tag is always given to a resource by a user. Therefore, the tag does not only provide information about the resource, but also about the user who annotated the resource. Through scanning previously used tags by the user, and also searching for users who use similar tags, a personalised tag recommendation can take place. So depending on the users' preferences, tags are recommended to the user depending on

- tags which were previously used by the user on similar resources,
- tags by similar users on the same resource,
- tags by similar users on similar resources and
- tags, which are used by users with common interests. This approach is also referred to as collaborative filtering.

Personalised recommendation systems can fail, when the users' interests and tagging behaviours change over time. Therefore, advanced personalised recommendation systems use a cyclic approach when looking at user behaviour [22].

Non-personalised Recommendation

This approach tends to recommend same tags for a certain item. Furthermore, tags are suggested to all users.

Content-based Approach: In content-based recommendation systems, textual sources are analysed in order to extract meaningful terms. Those terms can be unigrams or bigrams.

Analysing texts in those systems is done by various information retrieval methods. Those can for example include weighting terms by using TF/IDF scoring, analysing content through artificial neural networks which is trained on statistical information [20].

Collaborative Approach: When recommending tags using the collaborative approach, the system first looks for associated posts and extracts the tags assigned to this similar content. All tags are merged to one tag set, and thereafter reranked. Top ranked tags are then suggested to the user.

Personalised approaches outperform non-personalised approaches most of the time as they generally take more data into account [22].

4.2 Hashtag Recommendation

Hashtags are most popular on microblogging services. Those services and especially Twitter established the mainstream use of hashtags in social networking services.

4.2.1 Microblogging Systems

In Section 3.1 a brief overview about microblogging systems is given. Here, more details on this form of SNS are provided.

Twitter is a widely used and commonly known service for microblogging. Thus, properties of microblogging services are demonstrated with the help of this example. In contrast to a folksonomy, in a microblogging service a hashtag is not necessarily annotated to a certain piece of information. A hashtag is part of a tweet and can basically contain any form of information. Hashtags can represent the importance of a tweet to a particular group, or a particular topic. In microblogging services, most of the hashtags are user-generated. Due to this reason, hashtags contain user, topic or community specific language.

The importance of hashtags grows as the size of microblogging services grow. On average, 500 million tweets are posted on Twitter per day [34]. Hashtags are crucial to organise this amount of tweets. In a folksonomy, content is annotated for categorisation. The usage of a hashtag on the other hand depends on the content of the tweet. The content of a tweet is coined by the users' intentions and motives. Java et al. summerize some main user intentions on Twitter [15]:

Daily Chatter is the largest and most common topic Twitter. Users talk about their daily routines and what they are doing.

Conversations play a large role on Twitter. By using the -sign in order to address other users, Twitter makes it easy for users to speak directly to each other. Comments or replies to posts make up about one eighth of all tweets.

Sharing content (URLs): Around 13% of tweets contain URLs. As Tweets are restricted to 140 characters, URL shortening is a common strategy of gaining additional characters.

Reporting news and sharing latest events is common practice by a lot of Twitter users. Due to fast distribution and movement, recent news can be distributed easily through the microblogging service.

Furthermore, users on Twitter can be categorised due to their intentions. Java et al. propose following categories [15]:

Information Sources post statues on a regular intervals or infrequently. This type of users provide their relatively large number of followers with valuable and interesting content.

Friends are a very broad category. Nevertheless, most relationships on Twitter are of that nature. Interactions as casual conversations or daily chatter take place mostly amongst friends.

Information Seekers rather rarely post updates but follows other users on a regular basis.

A single user may have several intentions and serve several roles in microblogging services. The relationships in microblogging systems are more multifarious as they are in folksonomies. Therefore, it does not come as a surprise that hashtags are used in a different way than tags are.

4.2.2 Hashtag Recommendation in Microblogging Systems

As hashtags are not restricted in syntax, a very heterogenous group of hashtags occur on Twitter. Furthermore, users on Twitter use different hashtags to organise tweets wich belong to the same topic [25]. In order to increase search capabilities, some approaches on recommending appropriate hashtags have been made.

Kywe et al. ([17]) examined a dataset containing 44 million tweets by more than 150 000 Singapore Twitter users. They found, that only 8% of all tweets contain hashtags. Through automated hashtag recommendation, this number could increase and semantic value to tweets can be added. Furthermore they found, that the majority of hashtags have a short live span. Most of the hashtags used only occur in one tweet or are only used by one user.

In the following Subsections, a personalised recommendation approach as well as a non-personalised recommendation approach are presented.

Non-personalised Approach

Zangerle et al. ([25]) present one of the first approaches for Twitter hashtag recommendation. Their approach is to find hashtags for any tweet the user enters and recommend hashtags during the creation of the tweet. In order to analyse hashtagging behaviour of Twitter users, first a database containing 12 million tweets was set up.

The algorithm first searches for the most similar tweets compared to the users' tweet by using the previously built dataset. Then a set of hashtags from the similar tweets is retrieved. This set is then ranked and hashtag recommendation candidates are computed.

Similarity of tweets is determined by an adapted approach of term frequency - inverse document frequency method (tf/idf). The term frequency is the number of occurrences of one term in a tweet. the inverse document frequency provides information about the importance of a term within the whole set of documents (tweets) which are taken into consideration.

On order to rank the hashtag recommendation candidates, three ranking methods were suggested [25]:

Overall Popularity Rank which ranks hashtag recommendation candidates due to their popularity. Basically, the number of occurrences throughout the dataset determines the popularity of a hashtag.

Recommendation Popularity Rank which counts the occurrences of a

hashtag within the set of recommendation candidates. So the more often a hashtag occurs in similar tweets, the higher it is ranked.

Similarity Rank which is based on the similarity value of the users' tweet and the tweet containing the hashtag recommendation candidate. Similarity values are computed by the tf/idf method.

From these three methods, the Similarity Rank approach performed significantly better than the other two approaches.

Personalised Approach

As users on Twitter develop habits in using of the platform, also habits in using hashtags are developed [17]. British users may prefer british spelling also in their hashtags.

The personalised recommendation approach suggested by Kywe et al. [17] first accumulated a dataset containing 44 million tweets which were all posted in the area of Singapore. Their hashtag recommendation method selects hashtags from similar tweets as well as similar users. Those hashtag recommendation candidates are then ranked.

For finding similar tweets and users, the tf/idf method is used. Ranking hashtags is done by unifying the hashtags from the highest ranked similar users and the highest ranked similar tweets. Further ranking is done by comparing the frequencies of those hashtags. The more often it is used, the higher it will be ranked.

Kywe et al. found, that user preferences from few similar users significantly improve recommendation accuracy compared to non-personalised approaches.

4.3 Co-occurrance in Tag Recommendation

As pointed out in the Sections above, social systems which use tags and hashtags are not restricted in choice of language. Furthermore, no restriction applies to the users' choice and number of tags or hashtags. This freedom is a great advantage of those systems, as usage is easy and there are no limits for users' expression. At the same time, this feature marks the bottleneck in tag retrieval. This is due to different tags used for the same items, and other problems described in the Sections above such as language barriers or ambiguation. In order to overcome those problems, conceptually related tags have to be found. One obvious approach of finding similar tags is finding tags, which co-occur with one another.

Co-occurring tags are two tags which have been annotated to the same item. When talking about hashtags, co-occurring hashtags are two hashtags which have been used in the same entity e.g. a tweet.

A distinction is to be made between first- and second-order co-occurrence.

First-order co-occurrence looks at the co-occurrence of one tag with another tag. Basically, the number of two tags being annotated to the same item is counted. Second-order co-occurrence on the other hand takes the co-occurrence of one tag with all other tags into account. Here, for one tag all co-occurrences are counted. This means, that the number of all items which are annotated with both tags (the initial tag and the co-occurring tag) is considered. Through normalising this number by the number of all co-occurring tags, a distribution is found [23].

If co-occurrences are above chance, those are called significant. Furthermore, relationships between two co-occurring words are usually asymmetric. By using conditional probabilities from probability theory, asymmetric relations of words can be determined. Generally, the probability of an Event A under the condition of the event B is expressed by [16]

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{4.1}$$

In terms of co-occurrence, this can be pictured as follows. The probability that a word A occurs under the condition that another word B co-occurs, can be seen as a measure of the degree of association of word B with word A. This relation is asymmetric [16]. Section 5.2 will discuss the work done by Kubek et al. in more detail.

Chapter 5

Example Application

In this Chapter, an application is introduced which applies the concepts described in previous Chapters. This application aims to facilitate tagging for content creators in the field of music journalism. By semantically generating a tag set with the help of semantic web APIs, the content creator is automatically provided with an initial tag set. Tags from this initial tag set are then enhanced by social media. Co-occurring tags and hashtags from Soundcloud, Mixcloud and Twitter are ranked. Those tags are used to build an enhanced tag set from which the content creator can chose the most fitting ones.

5.1 Motivation

Content creation in Web 2.0 has become an increasingly harder challenge. First of all, the context of content creation is a very different one compared to traditional approaches. Content such as articles, tweets, posts or any form of short messages are exposed to a large audience in a split-second. Content consumers can become contributors by adding comments and thoughts to published articles. This can happen through conversations e.g. through "@"-mentionings on Twitter, or through comment functions on various platforms. Furthermore, content creators want to ensure that the content they are providing will be read by an according audience. Thus, strategies such as Search Engine Optimization (SEO), Tagging and publishing content on appropriate platforms are used. Choosing appropriate Hashtags when publishing the content increases ensures the appearance of content in the right context.

The concept of tagging is used in social media a lot. It has become normal for users to tag their content, and use hashtags in their posts. As a content creator one could benefit from the wide variety of tags and hashtags in social media.

In the field of music journalism, music-based social networks offer a wide

range of tags used and provided by artists and users alike. This is due to their popularity amongst artists. Artists use music based social media networks for song releases and promotion. As described in Section 3.3, content on music based social networks are often annotated with numerous tags.

In this application the services of Soundcloud and Mixcloud are used as those are popular amongst smaller and bigger artists alike. Furthermore, those networks are amongst the most fast moving networks. In 2013, 12 hours of content got posted on Soundcloud every minute [49].

But artists and bands do not exclusively use music-based social networks for distributing their content. Also, Twitter is a popular tool for finding, sharing and distributing music. The impact of Twitter usage in the field of new and trending music is relatively high compared to other social networking services. Twitter recently partnered with Billboard[1] to create Billboard real time charts [37]. Also in 2014, Twitter launched a service called Twitter-Music[2] which presents best new and trending music. Due to those reasons, Twitter can be considered as a powerful social media tool in music.

5.2 Related Work

Related works has been done in various fields. Co-occurrance in tag recommendation has been examined by Kubek et al. [16]. Also, the process of hashtag recommendation was investigated by several researchers such as Kywe et al. [17] for personalised hashtag recommendation and Zangerle et al. [25] for non-personalised hashtag recommendation. Lastly, the role of social tagging in music information retrieval was examined by Paul Lamere [18]:

Co-occurrance in tag recommendation: Kubek et al. use the conditional probability formula to determine the association between two co-occurring tags. They proposed an taxonomy-extraction algorithm based on this principle. In order to evaluate this algorithm, a taxonomy was extracted of Last.fm tags [16]. In this application, the enhanced tag set also relies on tag co-occurrance. Therefore, the approach by Kubek et al. will be discussed in more detail later in this section.

Non-personalised hashtag recommendation: Relying on similar tweets, the algorithm proposed by Zangerle et al. recommends hashtags, which have been used in similar tweets. Recommended hashtags are then ranked by different methods. The ranking method which performed the best, was based on similarity values between the initial tweet and all associated similar tweets [25]. This approach is also discussed in Subsection 4.2.2.

[1] http://www.billboard.com/
[2] https://music.twitter.com/

Social tagging in music information retrieval: Lambere investigates how social tags can be useful in the field of Music Information Retrieval. He concludes, that despite of the usual problems when dealing with social tags (noisy content, language barriers, etc. see Chapter 4), social tags can be used for genre, mood or instrumentation extraction of tracks [18].

Combining Social Music and Semantic Web: There has been some work done on combining social music and semantic web in order to enhance music recommender systems by Passant et. al. [47]. Social Networks are often used to suggest musical recommendations. Listening habits can be represented as on the last.fm exporter. Often, links between connections to artists, bands and interests are made. This undermines the importance of social networks in the field of music in general. Furthermore, Passant points out that interlinking datasets and the usage of combined data can be archived by contextualizing existing Web 2.0 data on the Semantic Web. Therefore, the Semantic Web benefits from previous web 2.0 and "Web of Documents" content [47].

The work done by Kubek et al. is essential for this application, as their approach is used for recommending co-occurring tags from social networks.

Based on the conditional probability theory, a formula can be derived. This formula gives insight on the association between a tag A and an co-occurring tag B. This is expressed by

$$Assn(A \to B) = \frac{|A \cap B|}{|A|}. \tag{5.1}$$

The formula describes the conditional relative frequency of tag B for the items annotated with tag A. The value calculated lies within a range between 0 and 1. Higher values indicate stronger associations between the tags.

5.3 Goals

By enhancing a semantically generated tag set through co-occurring tags and hashtags in social media, following theories shall be examined:

Finding trending tags. By determining if a currently trending tag is associated with one of the tags entered, reach may be increased by assigning this trending tag or using it when distributing the content on social media platforms.

Finding ambiguous tags. As described in Section 4.2, different hashtags are often used for the same topic. By including all hashtags which are used for one topic, the content may be exposed to a larger audience.

(E.g., using the hashtags #lykkeli #lykke #li #ll for annotating an article about the artist Lykke Li).

Determining context. Co-occurring tags of one submitted might or might not be out of context. The list of co-occurring tags, which is the basis for the enhanced tag set, should express the context of the submitted tag in social media.

First of all, content creators shall be provided with an automated tagging tool. The tagging tool should be easily accessible, and initiative to use. As their work takes place in an online environment, the application is designed as web application. Through this, the tool is not bound to any devices, or platforms.

By accessing the social component of those social media networks, a tag set shall be enhanced. The enhancement shall take advantage of the fast-moving community knowledge expressed by the tags and hashtags used. The enhanced tag set contains tags and hashtags used and generated by the community.

5.4 Technologies used

In order to stay independent in terms of technologies used in the individual platforms' APIs, this application is built with the help of state-of-the-art front-end web technologies. The application has been developed using JavaScript on the front-end as well as on the back-end side. In order to access various APIs which are crucial for the application, the application follows Representational State Transfer (REST) principles.

The nodejs server is a standalone server and is built on Chrome's JavaScript runtime. This server technology was chosen due to its easy scalability and the possibility of adding modules if the need appears. Used modules include for example restler[3], which makes REST API calls shorter and easier. Libraries such as jQuery[4] and underscore.js[5] were included as well as those libraries provide various supporting methods and functions.

The server runs on nodejs[6]. This ensures, that the application is scalable and lightweight. Furthermore, nodes uses an event-driven, non-blocking I/O model. The non-blocking approach is important to the application, as updates can be shown to the user while the client is still waiting for other calls to finish. On the server side, API calls and computation takes place. Due to this approach, API keys, data and computation methods are hidden from the front-end.

On the front-end side also JavaScript in combination with jQuery was

[3] https://github.com/danwrong/restler
[4] http://jquery.com/
[5] http://underscorejs.org/
[6] http://nodejs.org/

used. The framework Bootstrap[7] was used in order to speed up CSS development. Furthermore, a Bootstrap date picker plugin was added for displaying a date picker.

5.5 API Services used

The semantic web APIs provided by Zemanta, OpenCalais and Alchemy are used to establish an initial tag set. Those APIs, their advantages and drawbacks are discussed in Chapter 2.

Thereafter, social web APIs are accessed. The APIs are described and listed below.

5.5.1 Twitter Search API

Using the twitter search API, different parameters can be added to the query. Those include "result type" which can be set to "popular" or "recent". Here, Twitter already determines which tweets are more important than others regarding their amount of answers and retweets, and also the number of followers of the author. Furthermore, the API accepts a geocode which consist of the values latitude, longitude and radius. Then tweets submitted in those areas will be shown in the result set. Another important parameter which has been taken into account is the "created at" parameter. Setting this variable to a certain date, only tweets on or after that date will be returned.

The search API is limited to 100 tweets. Therefore, correlation values are higher in the mean, compared to topsy API, whose result limit is 1000 tweets. One benefit of receiving and processing only 100 tweets is, that results are shown to the user just after a few seconds of waiting. Therefore, the user can easily make several calls with different parameter and observe how the result sets differ. Also, for more recent tweets this APIs results are more convenient. So local trending topics can be accessed by the user easily by setting location and date values.

5.5.2 Topsy API

The service topsy has indexed all tweets since 2006. By calling the API, the 1000 most important tweets since then are returned. So using this API, a more representative result for an all-time correlation between tag A and tag B can be calculated. Topsy API also provides a parameter, which shows result from influential users only when set to true. This parameter is computed by taking retweets, public statements and other parameters into account. Through this method, topsy finds out who is listening to whom. In other words, the likelihood of people listening to a specific user is estimated.

[7] http://getbootstrap.com/

5.5.3 Soundcloud search API

When searching for sounds, those can be filtered by fields like “licence”, “duration” or “tag_list”. The latter is important for this application, as a set of B tags can be established easier by filtering sounds according to their tags.

The result set of Soundclouds API is restricted to a maximum value of 200 results per set. Due to this, several API calls are necessary in order to establish a data set to be analysed.

When searching for tracks, Soundcloud API offers parameters to be set. Those include a timestamp, but no geolocation values. However, if the parameter “order_by” is set to “hotness”, Soundcloud pre selects the result set according to the tracks’ performance on the social network. The performance is determined by likes, reposts and the play count. Also more recent tracks are considered more “hot” than older ones.

5.5.4 Mixcloud search API

Mixcloud API lets the developer search for either cloudcasts (tracks), users or tags. When searching for cloud casts, unfortunately no limitations concerning the published date or the popularity of the sound can be set. This API restricts the result set to 100 tracks per set. Though one can manually order the songs by date, a ranking of popularity proves to be difficult. Parameters such as “play_count”, “repost_count” or “listeners_count” could give some insight on the popularity of a track.

5.6 Application Structure

Before diving deeper into the systems’ architecture, the control flow and the user interface will be discussed.

5.6.1 User Interface and Control Flow

The interaction flow of the application is as follows:

1. The user posts the article into the text area.
2. A semantic web API has to be selected (Figure 5.1).
3. After submitting the text, the text is sent to the according semantic web API.
4. Tags or keywords extracted by the API are shown to the user. By assigning different opacity values to the background of the tag, the importance is indicated. Each item has a checkmark.
5. By checking one or more of the provided tags, the user indicates that this tag/s shall be used for enhancement (Figure 5.2). Furthermore, the user can add custom tags through an input field.

Figure 5.1: Selection field for semantic web APIs. Currently selected: Zemanta.

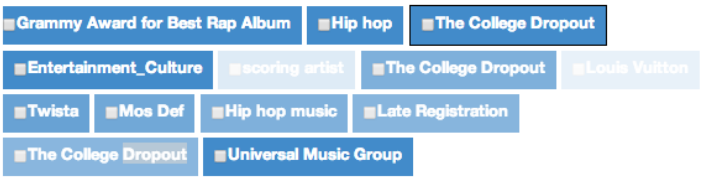


Figure 5.2: Initial Tagset generated by Alchemy API

6. Additionally, the user can enter a date. The date indicates the creation date of the social media items which are taken into account.
7. All the information is submitted, and co-occurring tags are extracted from each social media API.
8. Finally, the user will be provided with a table, which displays co-occurring tags and their association

5.6.2 System Documentation

The file index.js is entry point to the application. By calling node index.js the application is started locally. A server is created and listens to port 8080. So the application can be found at the url localhost:8080.

index.js creates a server, which listens to port 8080. Here, client requests are handled and all the communication with the client takes place. This class is the main controller of the application. Client events are handled and delegated to the according class. Events for the client are fired when the request is handled.

The files alchemyapi.js , zemantaapi.js and opencalaisapi.js are node modules, which provide the specified API for usage in node.js context. This approach was chosen to provide simple API calls for the file tagset.js. The class tagset.js provides the interface for calling the semantic web APIs. There also is a method for comparing results from different APIs, but this has no influence on the rest of the project whatsoever. Furthermore, tagset.js ensures, that the JSON objects returned by each API are normalized.

After a A-Tag set was created by using semantic web APIs the classes socialenricher.js, mixcloudenricher.js and musicenricher.js provide functionality for enriching the tag set. socialenricher.js provides the functionality for accessing twitter. Data from twitter can either be provided by twitter search

API or topsy API. By setting a boolean variable, twitter API or topsy API are used. First, the list of tags is submitted to the class. Then, a tag list of associated tags is extracted from the data provided by the API. Having received associated tags (a set of B-Tags). This set is given to the class unifyer.js in order to create a joint B-Tag Set containing the most relevant B-Tags from all three APIs. Same procedure is being used in the classes mixcloud.js for creating a B-Tag set out of tracks provided by mixcloud. The class musicenricher.js uses the same technique for extracting a B-Tag Set from soundcloud tracks.

As soon as all three APIs have submitted their B-Tag set to the class unifyer.js a priority set is created by the class. This set contains B-Tags, which were detected by at least two of the three APIs. The priority set is then given to the specific classes again (socialenricher, mixcloudEnricher and musicEnricher) for computing the value of co-occurrences of Tag A and Tag B. Their correlation with Tag A is derived by requesting entities (tweets or tracks) which show joint occurrences of tag a and tag b. After each finished computation for a B-Tag, the data is sent to the calling class (index.js) by using a callback. From index.js data is instantly sent to the client. This approach ensures that the user is able to receive the B-Tag as soon as the computation is finished.

On the frontend side, the class script.js is setting up the environment for communication with the server as well as reading data from the HTML and displaying data for the user. Furthermore, the location autocomplete functionality is implemented here. In order to receive city names and extract latitude and longitude values, an ajax call to the service geonames is called. This functionality can be used for a separate twitter search.

5.7 Evaluation

5.7.1 Results

Table 5.1 shows a search for the tag “Rock” with a result number of 1000. The time-span is set to one week, September 1st 2014 until September 7th 2014. In the result set the genres “funk”, “dance”, “jazz”, “hiphop”, “dubstep” and “electro” appear. The tag “funk” here shows higher association than other, less associated genres as “dubstep”. Furthermore, the appearance of locations should be remarked. A rather high Twitter association for “paris” (0.14) could indicate that an event associated with rock was taking place during that time period in Paris. Also, the locations “itally” and “philly” appear.

The tag “go” shows high association on Twitter as well. When examining the Tweets from those period of time, a common usage of those two hashtags can be found in prompting tweets, for example:

We need our Rock n Roll army! Request Save the World on YOUR #rock

Table 5.1: Tag A: "Rock"

Tag B	Mixcloud	SoundCloud	Topsy
party	0.02	0.08	0.08
booty	0.02	0.04	0.02
music	0.04	0.23	0.01
philly	0.02	0.1	0.02
funk	0.07	0.16	0.02
italy	0.03	0.14	0.02
go	0.05	0.15	0.818
dance	0.04	0.089	0.2
jazz	0.02	0.1	0.179
recording	0.04	0.08	0.1
hiphop	0.02	0.05	0.02
intro	0.07	0.04	0.02
dubstep	0.03	0.01	0.04
paris	0.07	0.02	0.14
will	0.07	0.04	0.02
die	0.07	0.17	0.02
electro	0.01	0.01	0.04

station. Ready? #GO![8]

When searching for genres, the associations with other genres are common. This is due to accessing the APIs of SoundCloud and Mixcloud. Tracks on those platforms are annotated with genres. In the given period of time, this association could indicate related events of both genres or music releases which cover both genres.

5.7.2 Bottlenecks

Unification of Semantic Web APIs

In the first steps of development, a unification of the semantic web APIs Zemanta, Alchemy and OpenCalais was planned. This proves to be difficult due to the following reasons:

- Those APIs were developed for different use cases. Therefore, different configuration variables are returned.
- The taxonomy used for representing entity types differs (AlchemyAPI schema[9], OpenCalais classes[10] and Zemanta entity types[11])

[8] Tweet by @adelitasway (https://twitter.com/adelitasway)
[9] http://www.alchemyapi.com/api/entity/types/
[10] http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/
[11] http://developer.zemanta.com/docs/entity_type/

- Measurement of importance of the entities differ. The relevance values of the APIs follow different approaches, therefore relevance values can not be compared with each other as they come out of the box.

An approach of unifying those (and other) semantic web APIs has been done by the NERD (Named Entity Recognition and Disambiguation) project[12]. In this projects, the classes of the different APIs were manually mapped. This was done by using their definitions and providing a best coverage of the principal axioms. Additionally, sub lasses were added to the newly created NERD ontology. The ontology was built on the retrieved classes. So the `nerd:City` displayed an equivalent to `alchemy:City` and `opencalais:City` while those are being more specific than `zemanta:location` [21].

The creators of the NERD project argue, that there is a need for a "gold standard" among entity extractors. This task is not easy to accomplish due to previously mentioned reasons.

As research by Bauer [1] shows, combination of semantic web APIs provides only little advantages. Through using multiple APIs, the keyword set can get incomprehensible and polluted by keywords with the same meaning but different spelling (e.g., Apple vs. Apple Inc.). This makes a error-free combination almost impossible.

Due to this reasons, a combination of keywords does not take place in this application. The user can decide which API shall be used. Through selection fields this decision can be made (see Figure 5.1).

Real-Time Web Application

In order to provide platform independence, the application has been developed as a web application. Advantages are, that the application can be accessed anywhere, from any device with a web browser. Furthermore, easy extension and comprehensibility is provided by using one programming language (JavaScript) in front-end as well as back-end side.

The application works in real time, meaning that the APIs are called as soon as the user clicks submit. This leads to waiting times, as the data has to be called and analysed after the user submitted. There is no dataset in the background which has been pre-analysed. Due to this, waiting periods can be up to 10 to 15 minutes, depending on the number of results demanded by the user.

API Restrictions and Dataset

As discussed in Section 5.5, the social media APIs used in this application provide different parameters, and therefore return according results. Also

[12] http://nerd.eurecom.fr/

restrictions on one API call vary throughout the APIs. Due to this, every API has to be handled separately.

Twitter search API provides the possibility of a geolocation, which can be used for local trending topics. The geolocation parameter is not provided by the APIs from SoundCloud and Mixcloud. Therefore, a combination of those APIs cannot take place when a geolocation parameter is added. Twitters results have to be considered separately if geolocation is set.

This lead to the decision of using Topsy API instead of Twitter API. Topsy provides large and reliable result sets. Twitter API has advantages when it comes to local analysis, but results are restricted to 1000 tweets.

Trade-offs between speed and the size of the result set had to be made. In order to analyse more data, larger result sets had to be generated.

The application offers the possibility of entering the number of desired results. Upper boundary of this number is 5000. As mentioned before, this leads to longer waiting periods.

Chapter 6

Conclusion

Through semantic annotation the web is enriched with metadata. Metadata can help indexing the web accessing information. This opens the possibility of numerous applications in this field. By using semantic web APIs, applications and mashups can be developed.

Combining semantic web APIs proved to be a difficult task. Furthermore, the combination of those APIs does not necessarily bring a lot of additional value to applications. New problems arise through combination. For example, confidence- and relevance values differ throughout the result sets provided by the APIs. Also, through combining entities from the APIs, inconsistencies can arise. Though one can argue, that in case of unavailability of one API, the user is still provided with results from the other APIs. Semantic extraction as a base for tag set generation is a good and reliable tool, which has been validated by this thesis. In order to enhance semantically generated tags, the social web can be taken into account.

When accessing tags and hashtags used in social media, the workings and usage of the media has to be taken into account. The social web offers a wide range of online networks, all of them serving different purposes. This thesis has shown, that tags and hashtags are used in different ways throughout the social online community. Therefore, the way of accessing content from social media is dependent of the characteristic of the social media stream, the characteristic of the social network and the tagging behaviour of the users.

Social media streams prove to be difficult to analyse by semantic web technologies. The content on social media streams differs from the information which is usually taken into account by semantic annotation tools. Content posted in online social media typically is short, noisy, temporal and always in a social context. Furthermore, the usage of different languages and slang language decreases machine-friendly readability.

But, posts on social media also contain tags or hashtags. Content in social media is therefore annotated already by the author of the post. By

exploiting the tags and hashtags already given by a user, semantic sense of social media posts can be extracted.

Hashtags in social media are used in different ways in different networks. On Twitter, often hashtags are user-generated and are used by the creator only. A majority of hashtags are used only over a short time period and by a specific group of users. Those factors make exploitation of hashtags more difficult.

On platforms as SoundCloud and Mixcloud, a large number of tags are used for an entity. Not all of those necessarily add value to the semantic meaning of the entity. This is due to tags such as “free”, “2014” or “good stuff”.

By analysing a large dataset, the drawbacks of tagging behaviour in those networks can be decreased. Analysing large datasets is a difficult task when trying to take the temporal factor into account. Social media is very fast moving, and most recent posts have to be taken into account. Due to this reason, the posts which are analysed in the example application are the most recent. By taking the most recent posts into account only, the relevance of the tags and hashtags extracted for the given point in time is increased. The liability of the results is decreased though, as the number of analysed posts can not be very high. Furthermore, when developing for a front-end real-time application, the waiting periods for the users should be kept as low as possible. This also requires to keep the number of posts to be analysed low.

The social web can be used as an extension for the semantic web in a lot of ways. When it comes to tagging, an exploitation of tags and hashtags used in different social networks helps enhancing tag sets. The enhancement shows that more colloquial terms are added to the tag set. This is a mirror of the language the internet, and more precisely the social networks which were taken into account, are speaking. When using social media for tag enhancement, one has to be aware that the tags and hashtags used are human-generated. Due to this, not always semantic value is added, but value for human readers.

References

Literature

[1] Andrea Bauer. “Analyse von Semantic Web-APIs und deren Methoden”. MA thesis. University of Applied Sciences, Campus Hagenberg, 2011 (cit. on pp. 13, 45).

[2] Tim Berners-Lee, Roy Thomas Fielding, and Larry Masinter. “RFC 3986: Uniform Resource Identifier (URI): Generic Syntax”. In: *Network Working Group* 66.3986 (2005), pp. 1–61 (cit. on p. 5).

[3] Tim Berners-Lee, James Hendler, Ora Lassila, et al. “The Semantic Web”. In: *Scientific American* 284.5 (2001), pp. 28–37 (cit. on pp. 4, 5, 8).

[4] Klaus Birkenbihl. *Standards für das Semantic Web*. Springer, 2006 (cit. on pp. 6, 7).

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. “Linked Data - The Story So Far”. In: *International Journal On Semantic Web and Information Systems* 5.3 (2009), pp. 1–22 (cit. on pp. 8, 10).

[6] Christian Bizer et al. “The RDF Book Mashup: From Web APIs to a Web of Data, 3rd Workshop on Scripting for the Semantic Web”. In: *CEUR Workshop Proceedings*. 2007 (cit. on p. 15).

[7] Kalina Bontcheva, Genevieve Gorrell, and Bridgette Wessels. “Social Media and Information Overload: Survey Results”. In: *CoRR (Computing Research Repository)* 1306.0813 (2013) (cit. on p. 19).

[8] Kalina Bontcheva and Dominic Paul Rout. “Making Sense of Social Media Streams through Semantics: A survey”. In: *Semantic Web* 5.5 (2014), pp. 373–403 (cit. on pp. 16, 19, 21).

[9] Tim Bray et al. “Extensible Markup Language (XML)”. In: *World Wide Web Consortium Recommendation REC-xml-19980210* (1998) (cit. on p. 6).

[10] Simon Carter, Wouter Weerkamp, and Manos Tsagkias. "Microblog Language Identification: Overcoming the Limitations of Short, Unedited and Idiomatic Text". In: *Language Resources and Evaluation* 47.1 (Mar. 2013), pp. 195–215 (cit. on p. 22).

[11] Fefie Dotsika. "Semantic APIs: Scaling Up Towards the Semantic Web". In: *International Journal of Information Management* 30.4 (Aug. 2010), pp. 335–342 (cit. on pp. 9, 10, 12).

[12] Bo Han and Timothy Baldwin. "Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 368–378 (cit. on p. 21).

[13] R. Irfan et al. "Survey on Social Networking Services". In: *Networks, IET* 2.4 (Dec. 2013), pp. 224–234 (cit. on pp. 18, 20).

[14] Bernard J. Jansen et al. "Twitter power: Tweets as electronic word of mouth". In: *Journal of the American Society for Information Science and Technology* 60.11 (2009), pp. 2169–2188 (cit. on p. 16).

[15] Akshay Java et al. "Why We Twitter: Understanding Microblogging Usage and Communities". In: *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*. Springer, Aug. 2007, pp. 56–65 (cit. on p. 32).

[16] Mario Kubek, Jürgen Nützel, and Frank Zimmermann. "Automatic Taxonomy Extraction through Mining Social Networks". In: *Proc. of the 8th International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods incorporating the 6th International ODRL Workshop, Namur Belgium*. 2010 (cit. on pp. 35, 37).

[17] Su Mon Kywe et al. "On Recommending Hashtags in Twitter Networks". In: *Proceedings of the 4th International Conference on Social Informatics*. SocInfo'12. Lausanne, Switzerland: Springer-Verlag, 2012, pp. 337–350 (cit. on pp. 33, 34, 37).

[18] Paul Lamere. "Social Tagging and Music Information Retrieval". In: *Journal of New Music Research* 37.2 (2008), pp. 101–114 (cit. on pp. 37, 38).

[19] Chi-Shiou Lin and Yi-Fan Chen. "The Influences of Online Cultural Capital on Social Tagging Behavior". In: *Journal of Library and Information Studies* 10.2 (2012), pp. 21–37 (cit. on pp. 23, 24).

[20] Cataldo Musto et al. "STaR: A Social Tag Recommender System". In: *Proceeding of ECML/PKDD 2009 Discovery Challenge Workshop*. Citeseer. 2009, pp. 215–227 (cit. on p. 31).

[21] Giuseppe Rizzo and Raphaël Troncy. "NERD: a framework for unifying named entity recognition and disambiguation extraction tools". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2012, pp. 73–76 (cit. on p. 45).

[22] M.M. Uddin, M.T. Hassan, and A Karim. "Personalized versus Non-Personalized Tag Recommendation: A Suitability Study on Three Social Networks". In: *Multitopic Conference (INMIC), 2011 IEEE 14th International*. Dec. 2011, pp. 56–61 (cit. on pp. 30, 31).

[23] C. Wartena, R. Brussee, and M. Wibbels. "Using Tag Co-occurrence for Recommendation". In: *Intelligent Systems Design and Applications, 2009. ISDA '09. Ninth International Conference on*. Nov. 2009, pp. 273–278 (cit. on p. 35).

[24] Jaewon Yang and Jure Leskovec. "Patterns of Temporal Variation in Online Media". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: ACM, 2011, pp. 177–186 (cit. on p. 21).

[25] Eva Zangerle, Wolfgang Gassler, and Günther Specht. "Recommending #-tags in Twitter". In: *Proceedings of the Workshop on Semantic Adaptive Social Web 2011 in connection with the 19th International Conference on User Modelling, Adaptation and Personalization, UMAP 2011*. Gerona, Spain: CEUR-WS.org, Vol. 730, 2011, pp. 67–78 (cit. on pp. 33, 37).

Online sources

[26] URL: http://blog.soundcloud.com/2013/10/28/hear-whats-happening-soundcloud-and-instagram/ (visited on 07/08/2014) (cit. on p. 19).

[30] URL: http://stackoverflow.com/questions/17114210/how-can-we-track-hashtags-with-the-new-facebook-hashtag-implementation (visited on 07/08/2014) (cit. on p. 24).

[32] URL: https://developers.google.com/+/api/latest/activities/search (visited on 07/08/2014) (cit. on p. 25).

[33] URL: https://dev.twitter.com/docs/api/streaming (visited on 07/08/2014) (cit. on p. 25).

[35] 2014. URL: http://www.google.com/trends/explore?hl=en-US#q=Twitter+Hashtag,+Facebook+Hashtag,+Instagram+Hashtag,+YouTube+Hashtag,+Google+Plus+Hashtag&date=1/2008+73m&cmpt=q (cit. on p. 25).

[36] Dave Beckett, ed. *RDF/XML Syntax Specification (Revised), W3C Recommendation*. 2004. URL: http://www.w3.org/TR/rdf-syntax-grammar/ (cit. on pp. 6, 7).

[37] *Billboard and Twitter Partner to Create 'Billboard Twitter Real-Time Charts'*. 2014. URL: http://www.billboard.com/articles/news/6028768/billboard-twitter-to-create-billboard-twitter-real-time-charts (cit. on p. 37).

[39] Rob DiCiuccio. *Entity Extraction & Content API Evaluation*. 2010. URL: http://blog.viewchange.org/2010/05/entity-extraction-content-api-evaluation (cit. on p. 17).

[38] Adam DuVander. *New API Billionaire Text Extractor Alchemy*. 2011. URL: http://www.programmableweb.com/news/new-api-billionaire-text-extractor-alchemy/2011/09/16 (cit. on p. 10).

[40] *Hashtags on Facebook Do Nothing To Help Additional Exposure*. 2014. URL: http://edgerankchecker.com/blog/2013/09/hashtags-on-facebook-do-nothing-to-help-additional-exposure/ (cit. on p. 25).

[41] Matthew Horridge et al. *A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools Edition 1.0*. Aug. 27, 2004. URL: http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf (cit. on pp. 7, 8).

[42] *Introducing RDF*. 2009. URL: http://www.linkeddatatools.com/introducing-rdf-part-2 (cit. on p. 7).

[43] Kevan Lee. *A Scientific Guide to Hashtags: How Many, Which Ones, and Where to Use Them*. 2014. URL: http://blog.bufferapp.com/a-scientific-guide-to-hashtags-which-ones-work-when-and-how-many (cit. on p. 25).

[44] Mark S. Luckie. *Best practices for journalists*. 2012. URL: https://blog.twitter.com/2012/best-practices-for-journalists (cit. on p. 25).

[45] Rebecca Murtagh. *The Role of #Hashtags in Social Media and Search*. 2013. URL: http://searchenginewatch.com/article/2305444/The-Role-of-Hashtags-in-Social-Media-and-Search (cit. on p. 22).

[46] Maximilian H. Nierhoff. *Hashtag Usage By Facebook Brand Pages Stagnates At 16%*. 2014. URL: https://www.quintly.com/blog/2014/01/facebook-hashtag-usage-stagnates/ (cit. on p. 25).

[27] Maximilian H. Nierhoff. *Research: Short Posts On Facebook, Twitter And Google+ Seem To Get More Interactions*. URL: http://www.quintly.com/blog/2013/12/short-posts-on-facebook-twitter-google-more-interactions/ (visited on 07/08/2014) (cit. on p. 21).

[47] Alexandre Passant and Yves Raimond. *Combining Social Music and Semantic Web for Music-related Recommender Systems*. 2008. URL: http://ceur-ws.org/Vol-405/paper3.pdf (cit. on p. 38).

[48] Emanuele Quintarelli. *Folksonomies: Power to the People*. June 2005. URL: http://www-dimat.unipv.it/biblio/isko/doc/folksonomies.htm (cit. on pp. 28, 30).

[49] *SoundCloud is 5!* 2013. URL: http://blog.soundcloud.com/2013/11/13/soundcloud-is-5/ (cit. on p. 37).

[34] *Twitter Usage Statistics*. URL: http://www.internetlivestats.com/twitter-statistics/ (visited on 08/07/2014) (cit. on p. 32).

[50] Thomas Vander Wal. *Folksonomy*. 2007. URL: http://vanderwal.net/folksonomy.html (visited on 07/08/2014) (cit. on pp. 27–29).

[29] *Wikipedia: Facebook*. URL: http://en.wikipedia.org/wiki/Facebook (visited on 07/08/2014) (cit. on p. 24).

[31] *Wikipedia: Google Plus*. URL: http://en.wikipedia.org/wiki/Google+ (visited on 07/08/2014) (cit. on p. 25).

[28] *Wikipedia: Hashtag*. URL: http://en.wikipedia.org/wiki/Hashtag (visited on 06/01/2014) (cit. on p. 23).