anacode®

Automated Market Research
October 2017

Anacode MarketMiner

Web-based Text Analytics for International Market Intelligence

Contents

1 Introduction

Text Analytics plays an important role in the "mix" of new, technology-based methods that are disrupting the market research domain. In combination with large quantities of up-to-date data from the largest database in the world, the World Wide Web, it can be used to perform continuous analysis and monitoring of customers, competitors and market trends. When compared to "traditional" market research, the advantages of Web-based Text Analytics are as follows:

- The Web offers a highly representative, natural and authentic data sample. By contrast, traditional market research normally works with smaller, artificially created data samples with a lower representativity.
- Whereas traditional market research has a static character, the Web allows for continuous monitoring and updating of insights.
- The automation of data collection and analysis allows to save cost, but also to process larger data quantities and thus achieve a higher coverage.

Text Analytics is situated at the intersection between the available data and the concepts and questions that are relevant for a business (Figure 1). One of the key challenges in building a successful Text Analytics application is defining, focussing and modelling this intersection. On the one hand, Web data is big, but not always relevant for a specific business domain. Besides, it contains large quantities of noise which affect both the quality and the speed of the analysis. For example, general "mood" reports in social networks hardly provide actionable information for brands. On the other hand, not all business questions and uncertainties can be solved using Web data, and those that can often require a reformulation into appropriate, technically feasible data queries. For instance, detecting the reason between low sales numbers for a specific product involves a multi-level sentiment analysis with gradual zoom-ins on the negative aspects of the product.

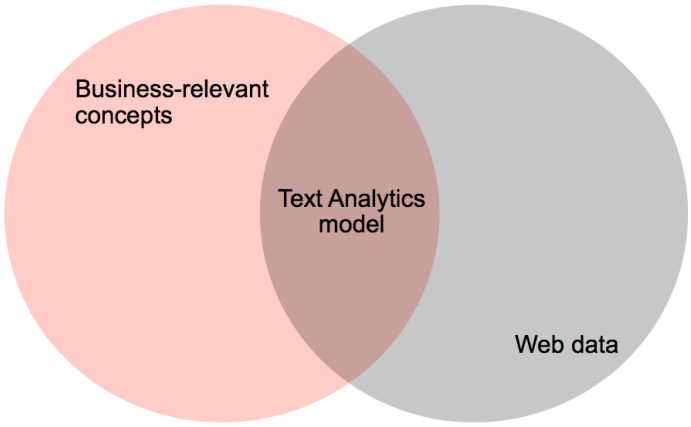

Figure 1: Text analytics at the intersection between business requirements and available data

This paper describes Anacode's end-to-end framework MarketMiner that is used to generate market insights from Web data. The framework covers all steps of the knowledge pipeline (cf. Figure 2):

1. **Data collection** collects large, diversified Web data into a unified database. Optionally, the data can also be combined with additional data provided by the customer.
2. **Linguistic analysis** (also Natural Language Processing) analyzes every text datapoint and provides its structured representation.

3. **Aggregation** works on a large quantity of structured datapoints and combines them into meaningful insights that can be visualized and used for further decision making.

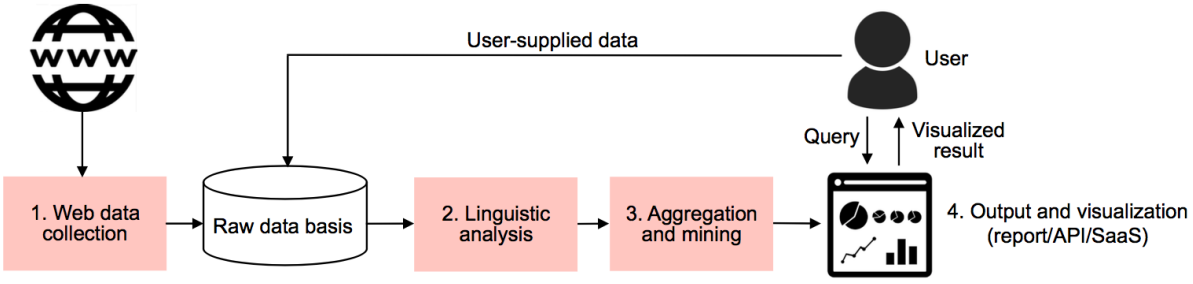

Figure 2: Pipeline of MarketMiner

The system is designed for scalability across different languages and thematic domains. Customization effort for specific languages and domains concentrates in the second step (linguistic analysis); it is kept to a minimum by the use of universal conceptual ontologies and the self-learning capacity of the algorithms. Currently, the technology is customized for Chinese and English. Work on a Russian component is in progress. It covers 30 major product and service industries, incl. Automotive, Finance, FMCG, Healthcare and Digital.

2 Data collection

2.1 A balanced data mix

Data – especially unstructured text data – is the crucial fuel of MarketMiner. On the one hand, it is the "raw material" from which MarketMiner generates insights and creates value for the user. On the other hand, new data continuously flows into the linguistic knowledge base and allows it to learn new words, linguistic patterns and entire semantic fields. As shown in Figure 2, there are two main sources of data: the Web and the company-internal data that can be provided by the user. The following Web data types are mostly used in Anacode's market analyses:

- **Social data**, which includes various subtypes such as social networks, product reviews and Q&A platforms, provides direct customer feedback and can be used to measure customer satisfaction and analyze market needs.
- **Professional data** includes specialized news and blogs which are written by specialized journalists. It has a high influence on the public opinion on specific topics and can be central for making purchasing decisions.
- **E-commerce data** provides information about the current market offering. It can provide detailed information on the properties, prices and purchasing modalities. Besides, e-commerce platforms often also include product reviews which allow to analyze direct customer feedback for each of the offerings.
- **Official data** mainly relates to websites of companies as well as official databases with company information. It can be used to mine factive business information, but also to identify softer aspects, for example the brand image and messaging of a specific brand.

Table 1 shows some of the Web resources that are monitored by MarketMiner.

Resource name - EN	Resource name - CN	URL	Professional	Forum	Reviews	Social
WeChat (public)	微信 (公众号)	http://www.wechat.com	✔	✖	✖	✔
Weibo	微博	https://www.weibo.com	✖	✖	✖	✔
Zhihu	知乎	www.zhihu.com	✖	✔	✖	✖
Sohu	搜狐	http://www.sohu.com	✖	✖	✖	✔
QQ	腾讯网	http://www.qq.com	✔	✔	✔	✖
Netease	网易	http://www.163.com	✔	✔	✔	✖
People	人民网	http://people.com.cn	✔	✖	✖	✖
Sina	新浪网	http://www.sina.com.cn	✔	✔	✔	✖

Table 1: Selection of monitored general Web resources

Beyond Web data, MarketMiner can also "digest" internal customer feedback provided by the client, such as call center logs, customer service emails and surveys in text form. As of now, most companies do not fully use all internally available data in their analytics. MarketMiner not only structures and extracts valuable insights from company-specific data sources, but also integrates them with external online data and thus provides a holistic picture of the entire feedback on the company and its products.

2.2 Web data collection

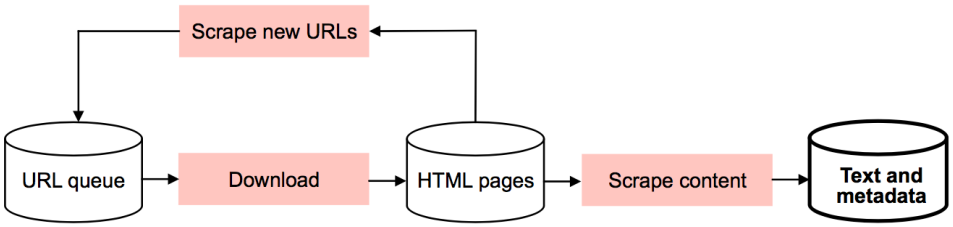

Figure 3: Data collection algorithm

MarketMiner uses a scalable data collection algorithm that can be flexibly used for a wide variety of data types and Web domains. The data collection proceeds in two steps: crawling, i. e. the download of the relevant webpages, and scraping, i. e. the extraction of meaningful content such as text and metadata. Figure 3 illustrates the workflow.

During crawling, all URLs identified on previously visited pages are saved to a URL frontier. At each new iteration, a new URL link is fetched from this frontier. The content found on the webpage of the new URL is added to the collected data, whereas the links collected from this page are again added to the frontier. The crawling algorithm can be customized for specific requirements. Additional filtering can be applied by imposing conditions on the URLs and/or on the crawled content. For example, it is possible to limit URLs to specific domains. For the content, linguistic filters such as keywords can be applied to make sure that the crawl stays thematically coherent and does not wander into irrelevant regions of the Web. This is especially relevant when collecting data on a very specific topic, such as a concrete brand or product model.

In the scraping step, MarketMiner extracts relevant content from the HTML and Javascript webpages that were downloaded. Three types of data are extracted:

- Text content
- Metadata, e. g. author, date, original source
- Influence metrics, e. g. numbers of likes and shares

Scraping is adapted for different types of webpages and can handle so-called "article" pages with a single block of content (e. g. *http://www.sohu.com/a/131626548_112589?loc=2&cate_id=998*) as well

as pages with a repetitive tabular structure, which is often the case for social media such as forums and product reviews (e. g. *http://car.bitauto.com/baoma5xichangzhoujuban/koubei/*).

The data collection is updated in regular time intervals. Most Web domains are updated several times per day using delta updating. Part of the collected data are available to registered API users (*https://api.anacode.de/web-data/*).

3 Linguistic analysis

Natural Language Processing (NLP) transforms unstructured data into a structured, table-like form which allows for further mathematical aggregation and data mining. NLP is an open-ended problem – for most tasks, it is impossible to get an analysis accuracy of 100%. The quantitative goal of an NLP system is to maximize precision and coverage while maintaining good performance and speed metrics. At Anacode, we believe that state-of-the-art NLP in the business context should also strive to satisfy the following conditions:

- Language is not a collection of words, but a complex system with numerous regularities, complex relations and alternative ways to express the same thing. NLP should always remain sensitive to the syntactic, semantic and pragmatic context of words and phrases.
- The implementation should allow a machine to build up its own intelligence from data by automatic learning, thus reducing human effort and increasing scalability.
- Whereas academic NLP work is often focussed on making algorithms work in "open" domains, real-world customer requirements are highly specific. The NLP component should learn and use domain knowledge and thus satisfy individual customer requirements.

Instead of betting either on a rule-based or a statistical approach, MarketMiner benefits from the joint potential of linguistic knowledge, state-of-the-art statistical algorithms and the power of Big Data. The following components and algorithms are used:

- **Lexical resources** such as dictionaries and ontologies are used to capture very common vocabulary, but also highly specific items such as the products of a brand or medical drugs in a given domain. A hierarchical organization in the form of an ontology allows to model relations between different types of concepts. For example, Figure 4 shows how concepts in the marketing domain are related in Anacode's ontology.
- **Grammatical knowledge** about the structure of words and phrases allows to bring a sentence into a tree-structured form which captures its underlying syntactic structure. On a "local" phrase level, chunking is applied to identify and analyse phrases of different types, such as nominal, adjectival and verbal phrases. On the sentence level, deep parsing can be used to construct full-fledged dependency tree representations which can be then used for semantic relation extraction.
- **Shallow machine learning** algorithms, such as Bayes classifiers, regression and Support Vector Machines, allow to learn lexical and linguistic regularities from labelled training data. They are especially helpful for text-level tasks with a large quantity of training data and a lower degree of difficulty, such as document-level sentiment analysis for a specific domain. In these cases, shallow models achieve both a high accuracy and good speed metrics.
- **Word vectors** are continuous representations of words in terms of their contexts. "Know a word by the company it keeps."[1] - it is a longstanding fact that a word is best described by the contexts in which it occurs. Word vectors are learned in an unsupervised fashion; the only input required is a large quantity of text data. Being algebraic constructs, they can be conveniently reused in further mathematical calculations.

[1] J. R. Firth (1957), Papers in Linguistics 19341951, London: Oxford University Press.

- **Deep neural networks** with multiple layers are inspired by neurons and their synapses in the human brain, which already indicates their high learning potential. In practice, neural networks allow for a more fine-grained feature representation than shallow machine learning approaches and thus have more predictive power. They are used for more challenging tasks such as fine-grained text categorization and concept extraction. In combination with word embeddings and large data quantities, neural networks allow to achieve the best accuracy metrics among the presently available machine learning algorithms.

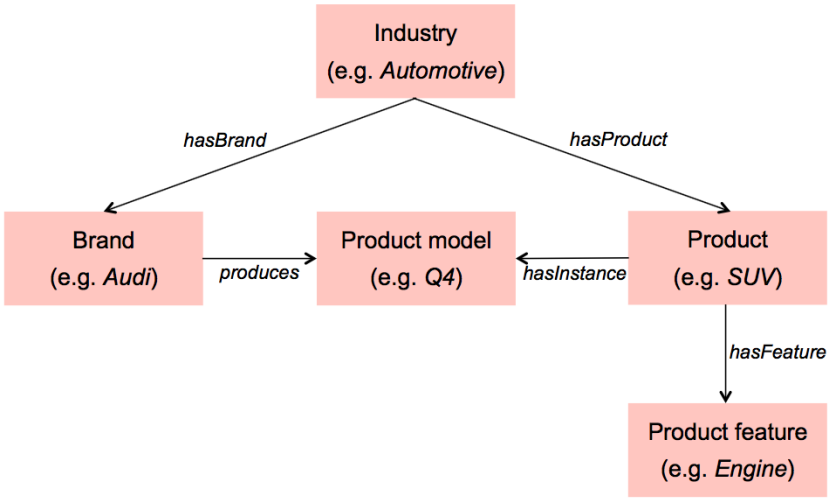

Figure 4: Excerpt of ontology concept types and relations

3.1 Algorithms

This section describes the algorithms that are used by MarketMiner to extract objective and subjective aspects of meaning. Table 3 shows the accuracy metrics for each of the production algorithms for Chinese and English.

Algorithm	Accuracy - Chinese	Accuracy - English
Text categorization	87.70%	88.11%
Concept extraction	86.32%	84.10%
Sentiment analysis (document-level)	93.12%	92.41%
Sentiment analysis (aspect-based)	84.69%	–

Table 2: Algorithm accuracy for Chinese and English

3.1.1 Text categorization

Text categorization allows to classify a text into one of several predefined categories. Anacode's standard API defines 30 top-level categories, such as Automotive, Cosmetics, Finance and Real Estate, which mostly correspond to major industries. Additionally, hierarchical classifiers are available that allow to classify texts at more fine-grained levels. Thus, the Automotive domain can be subclassified into more specific categories such as connectivity, e-mobility, automotive services etc.

3.1.2 Concept extraction

Concept extraction allows to identify marketing- and business-relevant concepts, such as products, companies, people, locations etc., from the provided text. Concepts can be manifested as single words (e.

Concept type	Description
organization	Companies, schools, associations etc., e.g. Central Bank, Arts School
location	Geographical locations, e.g. China, Moscow, Australia
person	People, professions etc., e.g. Teacher, Leo Tolstoy
time	Times and time expressions, e.g. April, Morning, 7:30
date	Dates in YYYY-MM-DD format, e.g. 2016-05-05
brand	Brands of goods and services, e.g. Apple Inc., TUI, Cartier
product	Physical products, e.g. Automobile, Mobile Phone
service	Non-material services, e.g. MBA, Concert, Surgery
product_model	Product models of specific brands, e.g. Golf, iPhone
product_feature	Product components and abstract attributes, e.g. Engine, Design, Comfort

Table 3: Extracted concept types

g. BMW) or phrases (Bayerische Motorenwerke). Each concept is associated with a specific type (cf. Table 4). Depending on their type, concepts can be annotated with additional semantic information; for example, brands also carry information about the industry they belong to, as well as the specific products and product models or series they produce (cf. Figure 4).

Generally, the implementation of concept extraction comes with several challenges:

- **Coverage**: the concepts of interest form open sets and are constantly updated with new items, such as newly created brands. Thus, it is impossible to provide an exhaustive list covering all possible concepts. To achieve a high coverage, the algorithm needs to be able to identify new concepts "on the fly" once they occur in the text.
- **Synonymy**: two linguistic expressions referring to the same concept (e. g. *car/automobile*, *engine/motor*) should be identified as synonyms denoting the same concept. This will allow to unify them in following aggregations.
- **Ambiguity**: many words and linguistic expressions are polysemous, which means that they can denote multiple concepts. For example, *bank* can refer both to the financial institution and the river bank. To maintain a high precision, these expressions need to be disambiguated based on their context.

MarketMiner uses a hybrid approach to concept extraction. First, rule-based concept matching is performed using the proprietary concept ontology, which currently contains ca. 40 thousands of the most common concepts. To increase the coverage of the algorithm, the second step is a statistical classifier, which uses word vectors to detect less known entities. This classifier is mainly based on a variation of the distributional similarity hypothesis: entities of a specific semantic class often occur in similar contexts. The new concepts detected by the classifier "on the fly" are manually checked and added to the ontology.

3.1.3 Sentiment analysis

Topics and concepts tells us *what* users are talking about. With sentiment analysis, we can learn *how* they are discussing it. Specifically, it detects the polarity of a text - whether it is positive or negative - as well as the intensity of the polarity. Sentiment can be understood from explicit sentiment expressions (e. g. *terrific*, *awful*, *to hate*), but also from more implicit signals such as events (e. g. *accident*, *launch*).

Sentiment analysis is a highly complex task which pertains to the whole spectrum of subjectivity in human language and cognition. The following challenges need to be addressed:

- **Mixed polarity**: a document can combine multiple polarities (e. g. *The service was very slow, but the food was delicious.*), each targeting a different entity. A document-level classifier should be able to correctly average these to a "neutral" polarity without getting biased into one direction.

- **Implicit opinions**: a sentence which is objective and neutral at the surface might still express a sentiment. This is frequently the case for sentences which describe events that humans automatically classify as positive or negative based on their knowledge of the world (e. g. *the engine produces a lot of noise*).

- **Sarcasm**: sarcasm (also irony) is a stylistic phenomenon whereby the author says the opposite of the truth in order to underline his dissatisfaction. For sentiment analysis, this means that a sentence which appears to be positive on the surface actually turns out to be very negative when considered in context. A sentiment classifier should be able to detect sarcastic sentences and reverse their surface polarity.

- **Domain dependence**: a word or statement can have different sentiments depending on the domain in which it is used. For example, *soft* might carry a negative connotation when describing a football player, but switch to positive once the person talks about pet toys.

- **Data type dependence**: the expression of sentiments and opinions is dependent on the data type. For instance, social data normally has a very high degree of subjectivity with lots of informal language, sarcasm etc. By contrast, official and professional data is more objective and less opinionated. For instance, news articles often don't express the personal opinion of the author, but it is still possible to derive the polarity from the events that are described (e. g. *hurricane* $\rightarrow$ negative, *victory* $\rightarrow$ positive).

- **Opinion spam**: especially in the domain of product reviews, opinion spam - fake reviews that are produced by a brand to promote its own products or demote competitor products - adds an additional dimension of noise. Spam filtering should be applied as a preprocessing step before the sentiment analysis and aggregation.

Since sentiment analysis is complex and multi-faceted, the first step is to "throw data at the problem". Appropriate and large training data sets are a necessary precondition for training a successful sentiment classifier. Human annotation is time-consuming and error-prone and can hardly produce data quantities large enough to be used for advanced algorithms such as deep neural networks. At Anacode, two automatic methods are mainly used for training data collection. First, raw metadata of certain data types - e. g. star ratings for product reviews - are used as proxies for the sentiment label. Second, distant supervision methods allow to generate training examples by bootstrapping from the initial linguistic knowledge base.

Besides, MarketMiner also uses hard-coded linguistic knowledge to capture the most frequent lexical and structural properties of linguistic sentiment. On the one hand, a sentiment lexicon contains items categorized by part-of-speech and annotated with their polarities and/or intensities. The following categories are covered:

- Evaluation adjectives, e. g. *good*, *horrible*

- Intensifiers and negators, e. g. *not*, *quite*

- Opinion verbs, e. g. *to appreciate*, *to dissatisfy*

- Nouns with an inherent evaluation, e. g. *advantage*, *obstacle*

On the other hand, a phrase structure grammar captures local structures, especially adjectival and verbal phrases that express opinions. Together with the concept extraction grammar focussed on nominal phrases, it is used for aspect-based sentiment analysis.

At the document level, specialized sentiment classifiers are available for each of the 30 industries that are covered by the text categorization algorithm. These can be used if the domain of a text is known in advance or had been detected using text categorization. If the domain of a text cannot be identified,

it is input to a generic sentiment classifier. This classifier is trained on a large quantity of data from different domains and thus performs with a high accuracy on open-domain data.

Aspect-based sentiment analysis identifies sentiment expressions and their "targets", i. e. the entities about which an opinion is expressed. Targeted entities are identified using the concept extraction algorithm. Sentiment expressions are identified based on the sentiment lexicon and grammar. A relation extraction algorithm is applied to match sentiment expressions to the correct targets. Optionally, information about the opinion holder can also be detected in those cases where the opinion holder is different from the author of the text. Figure 5 shows an example analysis and output for a Chinese sentence.

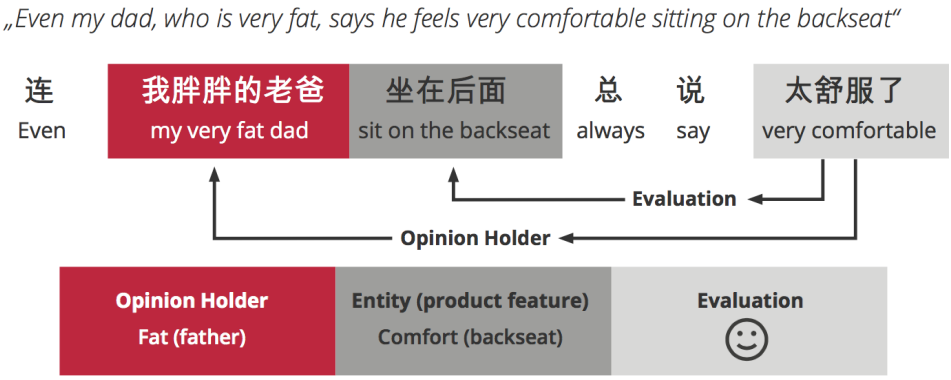

Figure 5: Example of aspect-based sentiment analysis

4 Aggregation and data mining

Whereas NLP provides a structured analysis for every datapoint in the data basis, the aggregation step combines the structured representations into a small number of distinct and meaningful analyses, ready for visualization and interpretation. Ultimately, the standard *wh*-questions can be answered based on the output of aggregation:

- **What are people talking about?** Using topic classification and concept extraction, MarketMiner extracts concepts that are often discussed online. These concepts can be aggregated based on frequency, relevance and co-occurrence.
- **How are they talking about it?** Sentiment and emotion analysis allows to analyze how the concepts are perceived by customers and other stakeholders. MarketMiner analyzes not only the polarities and emotions behind every concept, but also decomposes opinions into their different aspects and thus explains the "why" behind a specific sentiment value.
- **Who exactly is talking?** Author and user analysis allow to aggregate the data by authors and specific demographic characteristics. Additionally, key opinion leader ranking allow to identify those few online users that are particularly influential for the domain of interest.
- **When and where are they talking about it?** Metadata, especially information about the publish time of content and the location of the author, allows to build a spatial and temporal context around the identified insights which also can be used to filter down the data to a specific scope. Time series analysis is applied on the temporal information to detect emerging topic and sentiment trends.

Aggregation mainly works on the output of the NLP step. Besides, it uses the following additional information:

- **Metadata that are crawled additionally to the text data**. These can include information about the author, date, original source of the datapoint etc. Quantitative influence metrics, such as numbers of reads, shares and followers, allow to analyze the reach and influence of a piece of content.
- **Conceptual relations inferred by virtue of the ontology**:
  - **Hypernymy**, i. e. relations between super- and subclasses. For example, "electric vehicle" is a subclass of "new-energy vehicle" (NEV). Whenever electric vehicles are mentioned in the data, it is also inferred that the author is talking about NEVs.
  - **Meronymy**, i. e. relations between wholes and parts. This especially applies to physical product components. For example, upon talking about the nozzle, it is automatically inferred that the user is also talking about the engine.
  - **Custom relations** between concepts that are modelled in the ontology, e. g. relations between brands and their product series or models.