PRZEDSIĘBIORCZOŚĆ I ZARZĄDZANIE 2017
Wydawnictwo SAN | ISSN 2543-8190

Tom **XVIII** | Zeszyt **4** | Część **I** | ss. **85–101**

Anna Baj-Rogowska | Anna.Baj-Rogowska@zie.pg.gda.pl
Gdansk University of Technology

Agile Commerce in the Light of Text Mining

Abstract: The survey conducted for this study reveals that more than 84% of respondents have never encountered the term "agile commerce" and do not understand its meaning. At the same time, they are active participants of this strategy. Using digital channels as customers more often than ever before, they have already been included in the agile philosophy.

Based on the above, the purpose of the study is to analyse major text sets containing the "agile commerce" term, using the text data mining methods and tools. Data for analyses were sourced through creating an "agile commerce" query on websites with English as the content language. The texts retrieved in this way were used for building a corpus of documents. As a result of the corpus text exploration, words appearing most frequently in association with the agile commerce concept were separated and their most typical contexts were identified. The mining of unstructured data — the text mining process — revealed the essential meaning of the term being studied, while exploring knowledge from large information resources.

Key words: text mining, web content mining, agile commerce, mining unstructured data, knowledge exploration

Introduction

The Internet containing huge information resources is currently the most popular and at the same time the fastest source of retrieving various information. The wide access to huge resources makes it more and more difficult to distinguish important information from insignificant one and on that basis retrieve valuable knowledge. That knowledge is retrieving the right information needed to be used in the right time and suitable in the given situation. Speed of knowledge retrieving is the key factor.

Among the existing information resources in the Internet a significant group are text documents created in natural language. Extracting knowledge out of it is a labour and time consuming process. Thus the trend of developing the operation automation techniques assisting people in the process of retrieving knowledge from large information resources is visible. Computer systems equipped with the proper software are able to gather huge text resources and process them quickly and efficiently to retrieve the needed information. Thus the automated process of processing document created in natural language is called the explorative analysis of text data and generally known as text mining.

Text mining methods with implemented text set processing algorithms give the opportunity to search for information and retrieve documents according to defined criteria (topic, key notions), organise document structure and visualisation of the accumulated resources. They also allow to generate article summaries.

The study conducted for the needs of this paper[1] proved that over 84% of respondents has not come across and doesn't understand the term *agile commerce*. It became a reason to explore the Internet resources in the view to get to the essence of this term. The goal of the works was to perform the text mining analysis of large resources of non-structured data (in natural language) containing the term *agile commerce*. The aim of the research of the document corpus was to define what the *agile commerce* is. That made it possible to reach to the essence of the examined term by e.g. identification of the collocated keywords. The research also identified the detailed context and thematic areas in the Internet resources in which this term occurs the most often.

But before the text mining analysis process of the research term is carried out it needs to be established what the *agile commerce* strategy is. This issue will be discussed in the next point.

[1] The research was conducted at the turn of December 2016 and January 2017. The sample selection scheme was deliberate due to the chosen selection criteria (very good command of English). In the group of 45 people (academic staff and students) 19 advisers (over 42%) were selected.

Agile Commerce Strategy

As it was already mentioned, the research shows that over 84% of respondents have never encountered the term *agile commerce*. At the same time — as clients using the digital channels now much more than ever before — they are active participants of that strategy.

Polish e-commerce market is constantly and very dynamically developing. The clients value comfort, speed and attractive price offer of the products purchased in the web stores. In the situation when they are not buying online, very often they are searching the Internet for information regarding the purchasing process using computers and all kinds of mobile devices. As the Gemius report shows, 52% of Polish consumers uses the so-called multichannelling, namely start their shopping on one mobile device and finalise that process using another, e.g. PC [*E-Commerce w Polsce* 2015, pp. 132–134]. Digital channels bring new possibilities which business should use. Today it can be seen as obvious that the companies which provide better service in all channels — use *agile*, that is effective, smart, lively approach — can count on clients' loyalty and their future purchases. These observations led to a turn from the idea of Multichannel Commerce to Agile Commerce (also called Multichannel Commerce 2.0).

Multichannel commerce is a strategy of contacting with the clients by providing them with information through all available communication channels. While in *agile commerce*, in addition to providing the customer with as many places and events where they are dealing with a product or brand, the aim is a lively, active and flexible response to customer needs. National Research Council [1997, p. 308] defines *agile commerce* as integration of business with innovative IT whose aim is to (on the basis of electronic channels) build an agile network of connections among all business participants (e.g. clients, providers, etc.). This advanced — due to technology — network is to integrate information and efficiently distribute it to all its participants. Agility is believed today to be the next step in reengineering of business processes. Implementation of the agile approach changes the companies' paradigm of the realised processes, team work, communication and management [Metes, Gundry, Bradish 1998, p. 203]. This observation is reflected in the definition of *agile commerce* formulated by Brian Walker, the world-wide leader in multichannel commerce. It explains this term as "approach to trade which allows the business to optimise its staff, processes and technology so they best serve the clients in the area of all touchpoints" [Urbańska 2011]. Building the agile network helps the business to create successful alliances with partners and long-term relationships with clients.

Before the *agile* strategy occurred in commerce, it was successfully used in other fields. It is a term known from project management or software production based on iterative-incremental programming. The *agile* philosophy is perceived as very risky, but in practice in many areas it has proven its strength through its key features such as:

- fast reaction to changes,
- flexibility,
- strong cooperation with client.

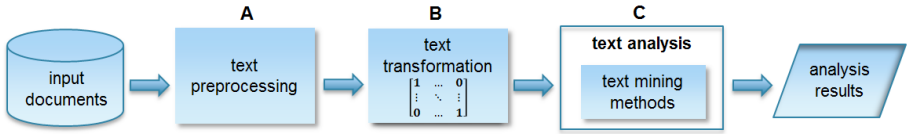
Technology development, constantly growing client's requirements, wide array of innovative products and large competition force the organisations to be more flexible and act effectively. Thus the agile approaches gain significance and they should be considered as a new way of thinking, adequate to today's reality.

Text mining — theoretical approach

Text data exploration defined as text mining (or text data mining) is concentrated on acquiring information from non-structured data through their processing. Its goal is to find new knowledge in the huge amount of text resources [Tuffery 2011, p. 627].

The process of text mining analysis of text documents is multi-stage and consists of phases which are outlined in the figure 1.

Figure 1. Stages of the process of text mining analysis of text documents



Source: own.

The non-structured text data retrieved for analysis — often recorded in various formats — are stored in one file. The data should be properly prepared and cleaned from irrelevant content so they can be structured (e.g. in vector model). For that purpose they undergo the processes described below.

A. Text preprocessing is called the stage of preprocessing the text file. This stage consists of the following sequence of operation [Miner, Elder, Fast et al. 2012, pp. 46–48]:

- **tokenisation** — transformation of words in the text in an orderly set of elements called tokens. Each token contains additional attributes like number allowing to maintain order of the tokens.
- **stop word list** — removing the words with low information value as well as other semantically insignificant words (like e.g. and, when, as well, but, with, etc.) which despite appearing in the text very often do not add to the content of the sentence, but only help to create the utterance.
- **stemming** — reducing the words to their basic form by finding their stems[2] which then are the representation of the word in the vector describing the document, e.g. the words *vector* and *vectors* will be treated equally since they have the common stem — *vector*.

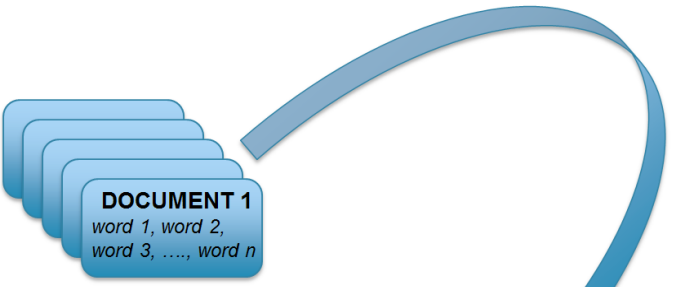
B. Text transformation — there are two approaches to mapping the way information is represented in the processed documents. The first one is based on the list of words in the document (vector space model), while in the other one the aim is to give a structured form to the information retrieved from the documents.

- **Vector Space Model**

  In this approach it is assumed that each document is represented by a vector whose each element inform about the number of occurrence of individual words [Berry 2013, pp. 3–4]. This model, most frequently used in practice, is colloquially called "bag of words". The words coming from individual documents are combined in one common list and they form the *matrix term-by-document frequency*. The matrix rows correspond with words in the documents and the columns represent documents. The idea of creating the *matrix term-by-document frequency* is shown in the figure 2.

[2] It needs to be emphasised that in English finding word stems is relatively easy. The classic method is Porter's algorithm [Tedd 2006, pp 38–40], which consists of five phases of word reduction. In the case of Polish, where inflection is irregular, finding stems is definitely more difficult. It is best to use lemmatization, that is finding the basic form of the word.

Figure 2. Example of representation with vector space model



	document 1	document 2	...	document N
word 1	1	0	...	1
word 2	0	1	...	0
...	...	...	...	0
word n	0	0	...	1

Source: own.

Words similarity is defined through analysis of similarity of corresponding rows in frequency matrix. Document similarity is defined by similarity of columns in the matrix. For classification of document features and their significance the TF-IDF weighing method is used [Chakraborty, Pagolu, Garla 2014, pp. 96–97]. To calculate it, the following weights need to be established:

- TF (*term frequency*) — is the number of word occurrence in the document;
- DF (*document frequency*) — is the number of documents where the word occurred;
- IDF (*inverse document frequency*) — represents the power of discrimination of texts by the given word, that is its goal is to calculate the significance of the word in the document on the basis of the number of documents containing this word.

$$IDF = \log(N/n_j) \tag{1}$$

where: N is the number of all documents,
n_j number of texts where the word occurs.

$$TF \cdot IDF = TF * IDF \tag{2}$$

By creating a matrix representing a set of texts mapped with a vector model, relations between the documents and the words contained therein are obtained. And so, based on the sum of the meanings of all the words of the document, one can get a good approximation of the substance of its content. Using statistical analysis of corpus content, it is possible to discover previously unknown relationships between words.

- **Representation in structured form**

In this approach, the transformation of the information contained in the text into a structured form can be accomplished by placing the words in the table records belonging to the relational databases. Then, using query engines (SQL), the data can be explored. A more recent approach is the creation of ontology, understood as a description of objects occurring in reality, together with a description of the relationship between them (so-called semantic relations). Such modelled system (objects — semantic relations) forms a semantic web. A good representation of its mapping are graphs.

C. Text mining methods and techniques [Weiss, Indurkhya, Zhang et al. 2010, pp. 7–14]:

- **document classification** — assigning new documents to one or more predefined classes, is very helpful when searching for or filtering information;
- **document clustering** — combining documents into groups (clusters) on the basis of their similarity, so that documents on one topic would fall in the same group;
- **summarisation** — generation of summaries; their simplest form is identification of the most important keywords with statistic methods; summaries can also be created on the basis of the most important sentences (taking into account their informational value) retrieved directly from the source text;
- **term clustering** — detection of dependencies between words occurring in a set of documents, identification of keywords;
- **visualisation** — presentation of search results and navigation in a large set of documents;
- **information retrieval, IR** — finding a subset of text documents referring to user's query in a large set of documents.

The text mining document analysis described above allows quick search in a large number of documents containing non-structured text data and retrieving words valid for analysis for which data mining algorithms are then applied. Practical implementation of text mining was presented in next point.

Text mining — practical approach

The purpose of the work was to implement a linguistic corpus analysis of the Internet resources related to the term *agile commerce*. This analysis is designed to identify the most frequently occurring words associated with the concept of agile commerce and to indicate in what context in the web resources this concept is most commonly found. Realisation of this task requires downloading and analysing text data from the Internet. WebCorp.org Live tool was used to obtain the most useful pages in this subject area. This is an online application that analyses the content of websites provided by its search engines (including Google and Bing). Through the query "*agile commerce*" on English-language pages[3] using the Bing (Cognitive) API, 22 text documents were generated from which the corpus was created.

For a text mining analysis a wide range of commercial applications (e.g. Statistica Text Miner, WordSmith, IBM Intelligent Miner for Text, SemioMaps, QDA Miner), and open source tools (such as TextSTAT, AntConc, RapidMiner, GATE) can be used. They offer varying levels of functionality ranging from the most basic statistical tasks to much more advanced abilities to build ontology and the use of advanced algorithms analysing the syntax of the text. All programs work well with analysing Western languages (some of them are even adapted to Japanese, Chinese or Portuguese), but their use for Polish texts poses difficulties resulting not only from encoding Polish characters, but also due to a different, complex syntax of Polish language. It can also be assumed that too small market for such a product is another impediment to the availability of text mining analysis programs for Polish. Due to the low availability of tools for the exploration of textual data in Polish, it was decided to download texts for analysis only from English-language sites.

ProSuite commercial software was used to analyse the corpus of documents. This is a program that provides advanced tools for complete and in-depth analysis of data, consisting of the following modules [cf. Provalis Research 2017]:

- QDA Miner (qualitative data analysis)
- WordStat (content analysis and text mining)
- SimStat (statistical analysis).

According to the procedure discussed in the previous section (fig. 1), all downloaded files were transformed into a corpus that was preprocessed. Non-structured data search was based on a model based on the multidimensional (vector) space created by all the words in the documents. Each corpus document should be interpreted as a vector consisting of *n* words whose coordinates determine the occurrence of the word in

[3] The choice of content from only English-language websites was dictated by the fact that the tools available to analyse texts were adapted to Western languages. A fuller explanation is provided later in this chapter.

the document. In order to be able to analyse the set, all text documents were transformed into an appropriate frequency matrix. A fragment of the data contained in the two documents D1 and D2 of the examined body is shown below. The frequency matrix (after the stop word list) created from them has assumed the form contained in table 1.

D1: … *The future of retail resides in agile commerce* …

D2: … *Agile Commerce is the fresh and new approach to commerce* …

Table 1. Fragment of frequency matrix taking into account binary representation

	…	*future*	*retail*	*reside*	*agile*	*commerce*	*fresh*	*new*	*approach*	…
D1	…	1	1	1	1	1	0	0	0	…
D2	…	0	0	0	1	1	1	1	1	…
…	…	…	…	…	…	…	…	…	…	…

Source: own.

By the use of stop-list the superfluous word and the ones little contributing to the analysis (prepositions, pronouns, conjunctions, etc.) were eliminated. Reducing word representations improves processing efficiency and also improves grouping results by removing information noise. The generated frequency matrix from the document corpus has become the basis for further analysis of textual data for the most common words in the text (sorting by frequency of occurrence). In figure 3 in the form of the so-called cloud a collection of words related to the agile commerce area that most often appeared in the examined corpus was visualised.

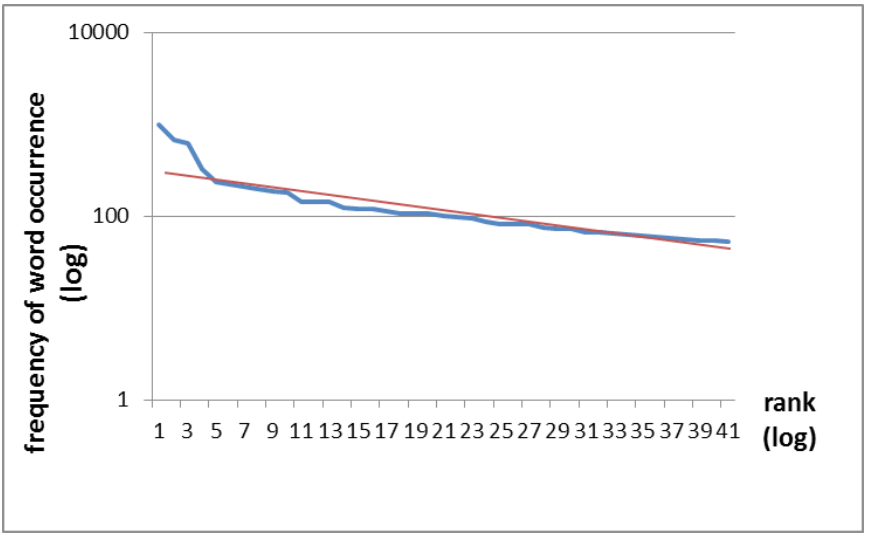
Figure 3. The most common words appearing in the examined texts in the area of *agile commerce*

Source: own.

Font size of each word reflects frequency of it occurrence, namely words occurring frequently in the corpus are represented by respectively larger font. The presented list contains unique words which apparently collocate well with the assumed research area and at the same time describe the substance of *agile commerce*. It does not include duplicated data, because in text mining, as a result of stemming, the words lose their inflected endings to better reflect the actual popularity of the word, measured by its occurrence in the examined text.

The graph with the logarithmic scale (figure 4) on the OX axis shows the rank of the occurrence of words in the examined document corpus, while on the OY axis the occurrence of a given word is shown. The straight line indicates the theoretical distribution corresponding to the Zipf's law. This law describes the frequency of use of particular words in any language [Berry 2013, p. 178]. It says that if for any text or group of texts a rank list is established (a list of words arranged in decreasing order of occurrence), the rank increases as the number in the list increases [Chakraborty, Pagolu, Garla 2014, p. 94]. The tendency described by Zipf's law is clearly visible in the examined material — the frequency is inversely proportional to rank.

Figure 4. Empirical and theoretical document corpus word distribution



Source: own.

The validity of words in the examined corpus was based on the TF-IDF measure. According to the formula (2) the weight of the word in the vector representing the given text is the product of TF and IDF weights. In this approach, the weight of a word

in a document is large, if the word occurs in it frequently. On the other hand, if the weight is decreasing, the more often the word appears in other documents. If a word has a high weight in a document, it means that it is well documented and will be useful when compared to other documents. These properties make it possible to identify the characteristic terms. Examples of such words in the examined body are given in table 2. It is clear here that the set of words for *agile commerce* clearly reflects its scope. Words with the highest weights make up a set of words well suited to the definition of the examined term.

Table 2. Words with the highest weights

term	*online*	*design*	*service*	*retail*	*customer*	*touchpoint*	*deliver*	*business*	*channel*	*technology*	*consumer*	*mobile*
term's weight	27,09	26,86	20,77	18,45	14,25	13,55	12,03	11,58	10,79	10,27	10,24	8,00

Source: own.

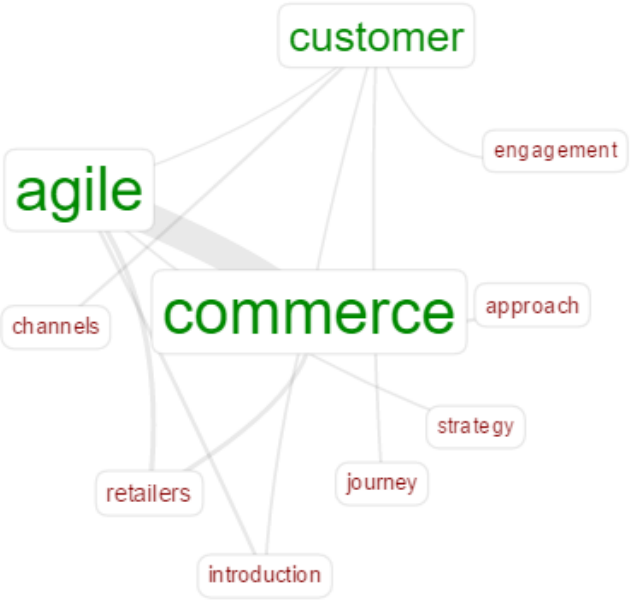
In the evaluation of the information retrieved from the documents a handy procedure may also be the analysis of relationships. It is a tool for identifying correctness in structured data, and the obtained results can serve as a basis for visualising existing relationships. The juxtaposition of word combinations (collocations) which in the corpus are accompanied by other words as constantly occurring patterns was shown in the figure 5. They are the results of KWIC research (*key word in context*). The words occurrence in the corpus was mapped with the size of the letters in the given word. Length and width of the lines among the words reflect how closely the two terms are collocated. The correlation between words *agile* and *commerce* (thick line) is strongly visible. The term *customer* plays in this topic a key role. It collocates with words: *agile*, *commerce*, *channels*, *journey*, *engagement*.

The collocation graph shows the words separated from the corpus are constantly repeating patterns and shows the relationships between them. It is at the same time a synthetic mapping of the components of the *agile commerce* strategy. In the examined material the following semantic relationships can be found:

- **agile** — **commerce**,
- **customer** — engagement, journey[4], channels,
- **agile** — strategy,
- **agile commerce** — approach, retailers.

[4] Journey and client are an effective agile strategy involving accompanying the consumer at every step of the purchase journey, urging him and subtly but regularly reminding of the offer.

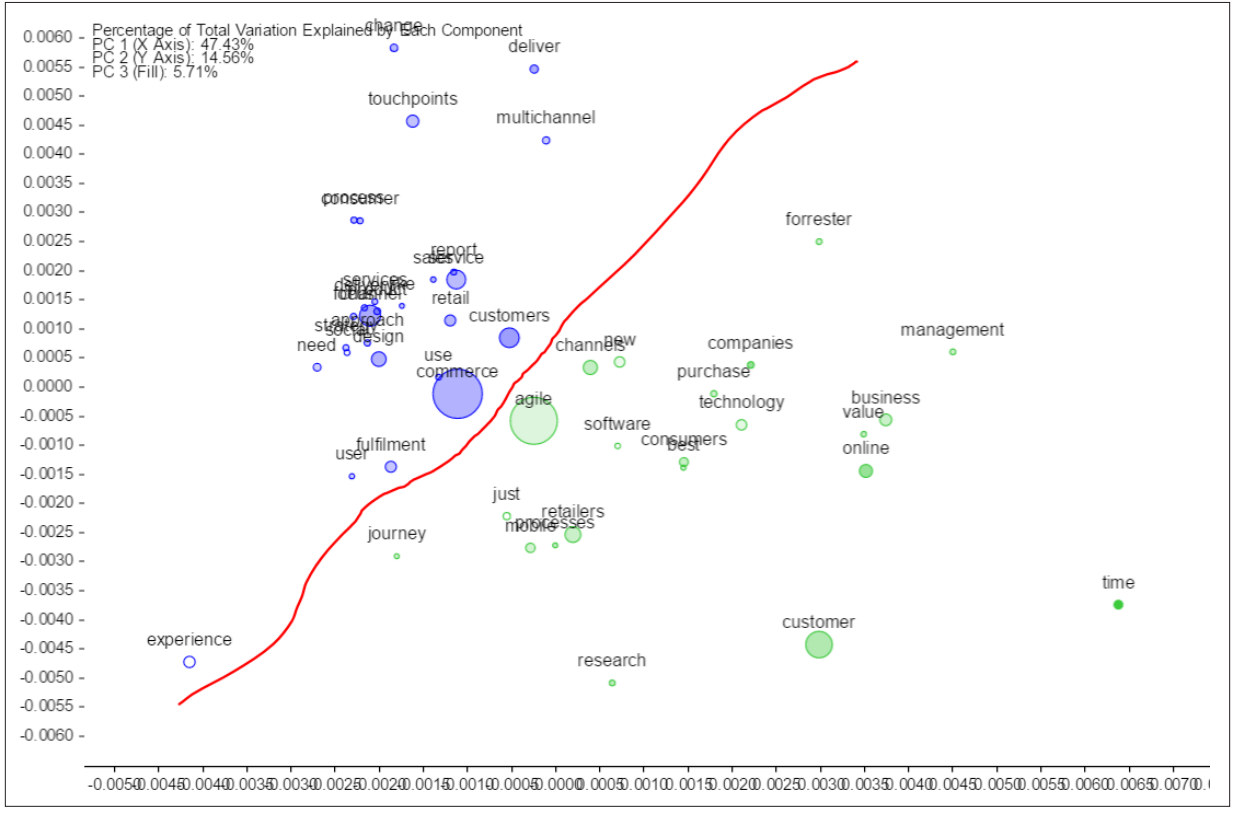
Figure 5. Collocation graph



Source: own.

A useful perspective in summarising data and detecting linear dependencies is the *Principal Component Analysis*, PCA) [more: Krawczyk 2012, pp. 98–99]. Its aim is to reduce the size of the data set and is often an indirect step in classification or cluster analysis. In the figure 6 we can see how the analysed words from document corpus are grouped. It is clearly visible that they don't create separate clusters (thus they are divided with a line). It indicates (once again) a strong positive correlation between the examined words (*agile – commerce*).

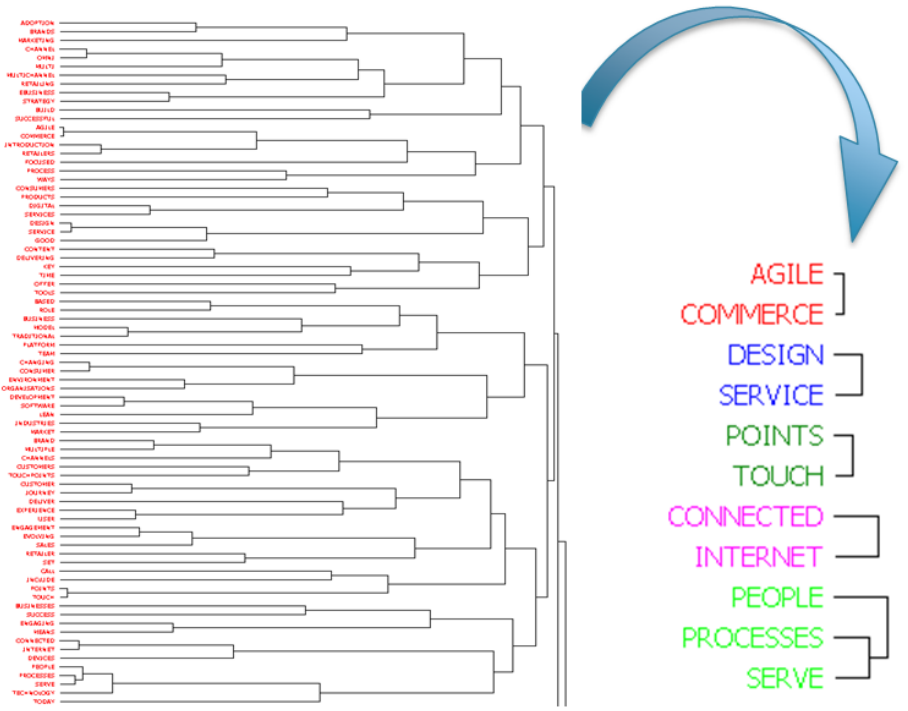
Figure 6. Principal Component Analysis



Source: own.

In the next step, the data set was subjected to a process called cluster analysis or clustering. This method is intended to identify disjoint subsets of words from the document corpus in relatively homogeneous classes. As the result of the agglomeration of hierarchical clustering the dendrogram presented in the figure 7 was obtained. Its construction was based on an agglomeration algorithm.

Figure 7. Cluster analysis in the dendrogram form



Source: own.

By increasing aggregation from the output level (left tree) we finally decided to get 5 clusters (dendrogram after transformation on the right in figure 7). The higher the level of aggregation, the less similarity between the elements of each class should be noted. On the basis of figure 7 it can be said that the examined corpus of documents contains the following content:

- developing/designing customer service,
- touch points with the customer,

- contact via the Internet,
- triads: people — processes — serve,
- phrases that overlap with the term *agile commerce*.

The use of clustering has allowed to isolate similar elements and combine them into homogeneous groups. It is obvious that the isolated clusters reflect all the components of the definition of the examined concept.

Text mining techniques also provide the ability to extract the main thematic threads most often appearing in the corpus of documents. The analysis identified a dozen or so topics that were eventually narrowed down to five core areas. Their synthetic form is shown in table 3.

Table 3. The main topics in the examined document corpus

No.	Topic	Keywords for the topic
1.	Online customer service	*support, product, services, platform, order, store, consumers, online, buy, systems, strategies, …*
2.	Components of agile commerce approach	*channel, serve, engagement, platforms, approach, social, business, technology, multiple, journey, consumer, devices, …*
3.	Shaping of purchasing process	*delivery, user, process, management, successful, lean, model, points, good, focus, buy, industry, success, …*
4.	Client — company communication via Internet	*connected, internet, products, things, today, calls, consumers, company, change, providing, design, devices, digital, …*
5.	Traditional vs digital purchasing model	*retail, purchase, control, set, traditional, buy, digital, touch, design, retailers, strategy, omni, …*

Source: own.

To sum up, the analyses conducted with the text mining tools and method have accurately defined the keywords and semantic collocations related to the term *agile commerce*. Their set and relations among the words are reflected in the definition of the examined term. The conducted analyses has also revealed the detailed context and thematic areas where the examined term occurs most frequently in the Internet resources.

Summary

Text documents as one of the most common form of storing and exchanging information from various areas of the world around us supply online databases exponentially. It results in growing need for systems allowing the automatic analysis of non-structured content, which will quickly and efficiently search, classify and summarise the information contained in the document. Thanks to statistical text analysis for occurrence of the certain keywords, word collocations, etc., these tools allow discovering knowledge through quick access to the essence of the searched content.

The corpus document study has shown that although it is possible to quickly and accurately retrieve information from a large set of text data, the search is limited to a certain group of items. Recognition of any information expressed by the natural language is a tremendously difficult task, as the natural language is characterized by complex and not always clear grammar and the complexity of semantic rules connecting the individual language constructs with their meanings. During text mining analysis of text documents, they are treated as mega collections of words in which the linguistic analysis of the processed text is omitted. The disadvantage of this solution is lower (because limited to a certain set) value of the obtained results. The undoubted advantage of text mining analysis is that it is fast and relatively simple (thanks to a wide array of IT tools) to discover knowledge from large collections of non-structured data.

Bibliography

Berry M.W. (red.) (2013), *Survey of Text Mining: Clustering, Classification, and Retrieval*, Springer Science+Business Media, Inc. NY, USA.

Chakraborty G., **Pagolu M.**, **Garla S.** (2014), *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*, SAS Institute Inc., North Carolina, USA.

E-Commerce w Polsce. Gemius dla e-Commerce Polska (2015), Gemius Polska [online], https://www.gemius.pl/files/reports/E-commerce-w-Polsce-2015.pdf, dostęp: 27.01.2017.

Krawczyk M. (red.) (2012), *Ekonomia eksperymentalna*, Wolters Kluwer Polska Sp. z o.o., Warszawa.

Metes G., **Gundry J.**, **Bradish P.** (1998), *Agile Networking: Competing Through the Internet and Intranets*, Prentice Hall PTR.

Miner G., **Elder J.**, **Fast A. i in.** (2012), *Practical Text Mining and Statistical Analysis for Non-structured Text Data Aplications*, Elsevier, UK.

National Research Council (1997), *Committee on Technology for Future Naval Forces, Commission on Physical Sciences, Mathematics, and Applications, Division on Engineering and Physical Sciences*, "Technology for the United States Navy and Marine Corps, 2000–2035 Becoming a 21st-Century Force", National Academy Press, Washington.

Provalis Research (2017), *ProSuite* [online], https://provalisresearch.com/products/qualitative-data-analysis-software/, dostęp: 12.01.2017.

Tedd L.A. (2006), *Program 1966-2006: Celebrating 40 Years of ICT in Libraries, Museums and Archives*, Emerald Group Publishing Limited, UK.

Tuffery S. (2011), *Data Mining and Statistics for Decision Making*, John Wiley & Sons Ltd., UK.

Urbańska N. (2011), *Czy nadchodzi koniec Multichannel Commerce?*, Ideas2Action [online], http://ideas2action.pl/2011/04/04/czy-nadchodzi-koniec-multichannel-commerce/, dostęp: 2.01.2017.

Weiss S.M., **Indurkhya N.**, **Zhang T. i in.** (2010), *Text Mining: Predictive Methods for Analyzing Unstructured Information*, Springer Science+Business Media Inc., NY.