GoldenBullet: Automated Classification of Product Data in E-commerce

Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klein, E. Schulten,
and D. Fensel

Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, NL, ying@cs.vu.nl

Abstract

Internet and Web technology starts to penetrate many aspects of our daily life. Its importance as a medium for business transactions will grow exponentially during the next years. In terms of the involved market volume the B2B area will hereby be the most interesting area. Also it will be the place, where the new technology will lead to drastic changes in established customer relationships and business models. B2B market places provide new kinds of services to their clients. Simple 1-1 connections are getting replaced by n-m relationships between customers and vendors. However, this new flexibility in electronic trading also generates serious challenges for the parties that want to realize it. The main problem here is caused by the heterogeneity of information descriptions used by vendors and customers. Intelligent solutions that help to mechanize the process of structuring, classifying, aligning, and personalizing are a key requisite for successfully overcoming the current bottlenecks of B2B electronic commerce. In this paper, we describe a system called ***GoldenBullet*** *that applies techniques from information retrieval and machine learning to the problem of product data classification. The system helps to mechanize an important and labor-intensive task of content management for B2B Ecommerce.*

1. Introduction

The World Wide Web (WWW) has drastically changed the on-line availability of information and the amount of electronically exchanged information. Meanwhile the computer has mutated from a device for computation into a entrance portal of large information volumes, communication, and business transactions (cf. [Fensel, 2001]). It starts to change the commercial relationships between suppliers and customers. Currently, a large fraction of the B2B transactions are still realized by traditional non-Internet networks, such as those conducted over EDI systems. In this traditional paradigm, direct 1-1 connections and mappings are programmed based on standards like EDIFACT (cf. [EDIFACT, 1999]). However, this traditional paradigm does not at all employ the full power of electronic commerce and it is quite likely that it will soon be out-ranged by more timely, Internet and web-based transaction types. Internet-based electronic commerce provides a much higher level of *flexibility* and *openness* that will help to optimize business relationships. Instead of implementing one link to each supplier, a supplier is linked to a large number of potential customers when linked to the market place.

However, preventing their customers from the bottleneck of facing exponential growth in the number of implemented business connections faces B2B market places with a serious problem. The have to deal with the problem of heterogeneity in *product*, *catalogue*, and *document* description standards of their customers. Effective and efficient management of different description styles become a key task for these market places.

Successful *content management* for B2B electronic commerce has to deal with various aspects: information extraction from rough sources, information classification to make product data maintainable and accessible, reclassification of product data, information personalization, and mappings between different information presentations [Fensel et al., 2001]. All of these sub-tasks are hampered by the lack of proper standards (or in other words by the inflation and non-consistency of arising pseudo-standards). The paper will focus on these challenges for content management and will discuss some potential solution paths.

The contents of the paper is organized as follows. In Section 2 we describe the overall content management problem that needs to be solved for effective E-commerce. Section 3 introduces our system **GoldenBullet** that applies information retrieval and machine learning techniques to one of the important sub-tasks of content management. **GoldenBullet** helps to mechanize the process of product classification. Section 4 provides an evaluation of our approach based on real-world data provided by B2B market places. Finally Section 5 provides conclusions and

Withold Abramowicz (ed.), Business Information Systems, Proceedings of BIS 2002, Poznan, Poland

discusses future directions.

2. Content Management in E-Commerce

B2B market places are an intermediate layer for business communications providing one serious advantages to their clients. They can communicate with a large number of customers based on one communication channel to the market place. The market places reduce the number of mappings to their user community from $n*m$ to $n+m$. However, in order to provide this service, they have to solve the significant mapping and normalization problem for their clients. A successful market place has to deal with various aspects. It has to integrate with various hardware and software platforms and has to provide a common protocol for information exchange. However, the real problem is the heterogeneity and openness of the exchanged content. Therefore, *content management* is one of the real challenges in successful B2B electronic commerce. It tackles with a number of serious problems [Fensel et al., 2001]:

1 Product descriptions are unstructured.

2 Product descriptions are unclassified.

3 Product descriptions must be classified and described in various dimensions because no standard product classifications exist.

Product descriptions must be structured. Suppliers have product catalogues that describe their products to their potential clients. This information should be made on-line available by a B2B market place. One could think that this may be a simple task because most product catalogues already exist electronically. However, these product catalogues are designed for the human reader. Extracting the actual product information and storing it in a structured format is therefore mainly a manual task. A content management solution provider like Content Europe[1] has several hundred employees working in content factories to manually structure the product information. In the worst case, they take printed copies of the product catalogues as input.

Product descriptions must be classified. At this stage in the content management process we can assume that our product information is structured in a tabular way. Each product corresponds to an entry in a table where the columns reflect the different attributes of a product. Similar products are group together in the same table. Each supplier uses different structures and vocabularies to describe its products. This may not cause a problem for a 1-1 relationship where the buyer may get used to the private terminology of his supplier. B2B market places that enable *n-m* commerce cannot rely on such an assumption. They must classify all products according to a standard classification schema that help buyers and suppliers in communicating their product information. A widely used classification schema in the US is UNSPSC[2] (for details about UNSPSC, please see next section). Again it is a difficult and mainly manual task to classify the products according to a classification schema like UNSPSC. It requires domain expertise and knowledge about the product domain.

Product descriptions must be re-classified. Bottlenecks in exchanging information have led to a plethora of different standards that should improve the situation. However, usually there are two problems. First, there are too many “standards”, i.e., none of them is an actual standard. Second, mostly, standards lack important features for various application problems. Not surprisingly, both problems appear also in B2B electronic commerce. UNSPSC is a typical example for a *horizontal* standard that covers all possible product domain, however, is not very detailed in any domain. Another example for such a standard is the *Universal Content Extended Classification (UCEC)*[3]. It takes UNSPSC as a starting point and refines it by attributes. Rosetta Net[4] is an example for a *vertical* standard describing products of the hardware and software industry in detail. Vertical standards describe a certain product domain in more detail than common horizontal ones. More examples for such “standards” can be found in [Fensel, 2001].

In the reminder of the paper we focus on one of these sub-tasks. We will describe our solution we developed for product classification. However, we would also like to mention that we are currently evaluating similar techniques for product data structuring and re-classification.

3. GoldenBullet

Finding the right place for a product description in a standard classification system such as UNSPSC is not at all a trivial task. Each product must be mapped to the

[1] http://www.contenteurope.com

[2] http://www.un-spsc.net and http://www.unspsc.org.

[3] http://www.ucec.org

[4] http://www.rosettanet.org/

***XX* Segment**
The logical aggregation of families for analytical purpose
***XX* Family**
A commonly recognized group of inter-related commodity categories
***XX* Class**
A group of commodities sharing a common use or function
***XX* Commodity**
A group of substitutable products or services
Business Type
The function performed in support of the commodity. This value is vendor specific and does not affect the base code for each commodity.

Figure 1. The layered structure of UNSPSC.

corresponding product category in UNSPSC to create the product catalog. Product classification schemes contain huge number of categories with far from sufficient definitions (e.g. over 12,000 classes for UNSPSC) and millions of products must be classified according to them. This requires tremendous labor effort and the product classification stage takes altogether up to 25% of the time spent for content management. Because product classification is that expensive, complicated, time-consuming and error-prone. Content Management needs support in automation of the product classification process and automatic creation of product classification rules.

GoldenBullet is a software environment targeted to support product classification according to certain content standards. It is currently designed to automatically classify the products, based on their original descriptions and existent classifications standards (such as UNSPSC). It integrates different classification algorithms from the information retrieval and machine learning areas and some natural language processing techniques to pre-process data and index UNSPSC.

3.1. UNSPSC

The Universal Standard Products and Services Classification (UNSPSC) is an open global coding system that classifies products and services. It was first developed by Dun & Bradstreet and the United Nations Development Program. It is now maintained by the Electronic Commerce Code Management Association (ECCMA) which is a not-profit membership organization. The UNSPSC code covers almost any product or service that can be bought or sold, which includes 12,000 codes covering 54 industry segments from electronics to chemical, to medical, to educational services, to automotive to fabrications, etc. The UNSPSC is heavily deployed around the world in the electronic catalogs, search engines, procurement application systems and accounting systems. It is a 10 digit hierarchical code that consists of 5 levels (see Figure 1).

3.2. Overall GoldenBullet Functionality

GoldenBullet as a software environment provides the following functions to fully achieve semi-automatic or automatic product classification: Data input and export facilities; text processing techniques; classification of product data; and learning and enrichment of product classification information (see Figure 2).

3.2.1 Data Input, Output, and Validation

A wrapper factory gathers various wrappers to convert raw data description from external formats to internal format, and final results to preferable output format or

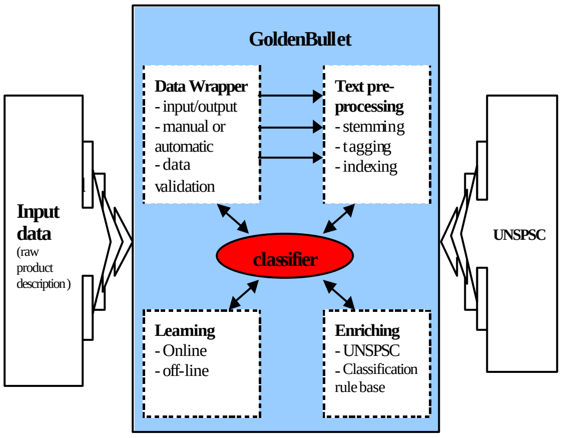
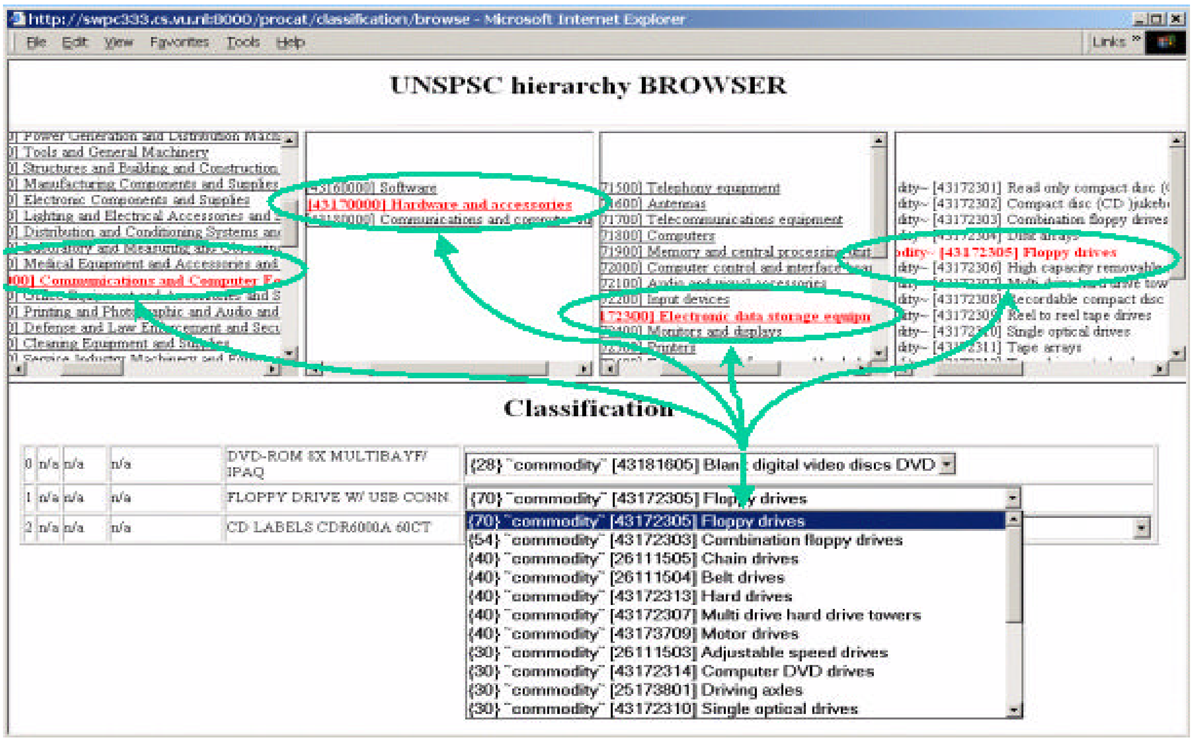


Figure 2. Overview on **GoldenBullet**.


http://swpc333.cs.vu.nl:8000/procat/classification/browse - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links »

UNSPSC hierarchy BROWSER

0] Power Generation and Distribution Machi
0] Tools and General Machinery
0] Structures and Building and Construction
0] Manufacturing Components and Supplies
0] Electronic Components and Supplies
0] Lighting and Electrical Accessories and
0] Distribution and Conditioning Systems and
0] Laboratory and Measuring and Observing
0] Medical Equipment and Accessories and
00] Communications and Computer Eq
0] Office Equipment and Accessories and S
0] Printing and Photographic and Audio and
0] Defence and Law Enforcement and Secu
0] Cleaning Equipment and Supplies
0] Service Industry Machinery and Equi

[43160000] Software
[43170000] Hardware and accessories
[43180000] Communications and computer ac

71500] Telephony equipment
71600] Antennas
71700] Telecommunications equipment
71800] Computers
71900] Memory and central processing unit
72000] Computer control and interface equi
72100] Audio and visual accessories
72200] Input devices
172300] Electronic data storage equipm
72400] Monitors and displays
72500] Printers

dity~ [43172301] Read only compact disc (
dity~ [43172302] Compact disc (CD)jukeb
dity~ [43172303] Combination floppy drives
dity~ [43172304] Disk arrays
odity~ [43172305] Floppy drives
dity~ [43172306] High capacity removable
dity~ [43172307] Multi drive hard drive tow
dity~ [43172308] Recordable compact disc
dity~ [43172309] Reel to reel tape drives
dity~ [43172310] Single optical drives
dity~ [43172311] Tape arrays

Classification

0	n/a	n/a	n/a	DVD-ROM 8X MULTIBAY F/ IPAQ	{28} ~commodity~ [43181605] Blank digital video discs DVD
1	n/a	n/a	n/a	FLOPPY DRIVE W/ USB CONN	{70} ~commodity~ [43172305] Floppy drives
2	n/a	n/a	n/a	CD LABELS CDR6000A 60CT	

{70} ~commodity~ [43172305] Floppy drives
{54} ~commodity~ [43172303] Combination floppy drives
{40} ~commodity~ [26111505] Chain drives
{40} ~commodity~ [26111504] Belt drives
{40} ~commodity~ [43172313] Hard drives
{40} ~commodity~ [43172307] Multi drive hard drive towers
{40} ~commodity~ [43173709] Motor drives
{30} ~commodity~ [26111503] Adjustable speed drives
{30} ~commodity~ [43172314] Computer DVD drives
{30} ~commodity~ [25173801] Driving axles
{30} ~commodity~ [43172310] Single optical drives

Figure 3. A screenshot of **GoldenBullet**.

user-designed formats. Besides the automatic importing and exporting data, **GoldenBullet** also provides the editor for manually inputting data, which suits well for small and medium vendors.

3.2.2 Text pre-processing

The validated product data will be pre-processed before the classification has been performed. Some of the Natural Language Processing algorithms have been implemented into **GoldenBullet**. The product data will be stemmed (grouping different words with the same stems) and tagged (extracting noun-phrases). Furthermore, UNSPSC is also being pre-processed (stemmed and tagged) to make sure that noisy words or information have been screened. A stop word list has been generated, updated and extended during the whole process. Currently, **GoldenBullet** can handle English and French product data.

3.2.3 Product Classification

Figure 3 shows the user interface of the classifier. The imported UNSPSC is browsable from the screen, which directs the end user to the right location of UNSPSC. The classifier classifies the pre-processed product data and proposes the ranked solutions based on various weighting algorithms. The end user can pull down the proposed list and make the final choice. But when he highlights one of the proposed solutions, the above UNSPSC browse window will show the exact location of it in UNSPSC with the details of each level.

Performing the classification task is viewed as an information retrieval problem (see [Ribiero-Neto & Baeza-Yates, 1999] for an introduction to the field). The problem of finding the right class is viewed as the problem to find the right document as an answer to a query:

- A product description is viewed as a query and UNSPSC is viewed as a document collection.
- Each of the commodities in UNSPSC is treated as a document, where each commodity description forms the text of the document.
- Assigning a proper category for a product is achieved via retrieving a correspondent UNSPSC commodity description.

The performance of such an approach is rather low (see the next sub-section for more details). Directly using UNSPSC as document collection fails in this respect because the class descriptions are very short (i.e., we deal with very short documents) and the product descriptions are often very short too and use very specific vocabulary that cannot directly be matched with more generic terms in UNSPSC. Evaluation with real world data showed that less than 1% of actual product data could classified correctly which such a naïve approach. Therefore, we employed various strategies to achieve a more reasonable and workable result that are described in the next subsection. Basically we employed different retrieval strategies and we made use of large volumes of manually classified data to improve the performance.

3.3. The heart of GoldenBullet: Intelligence and Knowledge

The essence of **GoldenBullet** is its ability to automatically classify product descriptions. This sub-section will present our classification approaches and show how they can make use of pre-classified data as background knowledge. An evaluation of the approaches in given in section 4.

3.3.1 The Vector Space Model (VSM)

A standard method in Information Retrieval is the well-known *Vector space model (VSM)*. Salton's Vector Space Model (cf. [Salton et al., 1975]). It uses the word vector to represent document and user query, then applies the cosine similarity formula to calculate the similarity between the document and query so as to retrieve the most relevant document to user's query. The same model has been applied in text categorization. [Gomez-Hidalgo & Rodriguez, 1997] used Salton's vector space model to represent document (in our case product description) and existing categories (e.g. in our case UNSPSC). Then the category (UNSPSC) can be assigned to a document (product) when the cosine similarity between them exceeds a certain threshold. The basic idea is to represent each document as a vector of certain weighted word frequencies.

VSM is adopted by us to find the match between UNSPSC commodities and product descriptions. In the following, we will describe two strategies how VSM can make use of pre-classified examples in two ways. Both strategies treat an unclassified product description as a query, however, differ in what they us as a document collection.

1 The first version takes each commodity as a document. The examples are used to enrich the commodity description. Basically we extract words from pre-classified product data and add them to the word list describing the commodity.

2 The second version takes each pre-classified product description as a document. We use VSM to retrieve the instance fitting best to a newly product description and infer the UNSPSC code of the latter from the known UNSPSC code of the former.

We will describe pros and cons of both approaches in section 4.

3.3.2 K-nearest neighbor

Another instance-based classifier we implemented is based on the k-Nearest Neighbor method **KNN**. Again, the algorithm uses the set of pre-classified examples directly to classify an example. The algorithm passes the whole set of training examples and searches for the most similar one, and then assigns the class to the new example, equal to the class of the most similar one. **KNN** is computationally expensive and requires lots of memory to operate depending on the number of pre-classified examples. Again we can distinguish two modes in regard to whether the algorithm works directly on the pre-classified product data or on enriched class descriptions.

3.3.3 Naïve-Bayes classifier

The final paradigm we employed is the machine learning paradigm (see [Mitchell, 1997] for a good introduction and analysis of the field). This paradigm assumes existence of a set of (manually) pre-classified products, which is called a training set, and a set of product descriptions to be classified by the systems, which is called a test set. The Naïve-Bayes classifier **NB** [Burges, 1998] uses Bayes theorem to predict the probability of each possible class, given a product description and the set of training pre-classified examples as input.

The classifier assigns the commodity, which has the highest probability of being correct. Naïve-Bayes is a standard text classification algorithm, with a long successful application history.[5]

3.3.4 Hierarchical Classification

UNSPSC provides a hierarchy of four levels for classifying products: *Segment*, *Family*, *Class*, and *Commodity*. Therefore, it is quite natural to employ a hierarchical classification approach for our task. In addition, we made the experience that lots of pre-classified products we received for our evaluations we report in the next section are not classified up to the

[5]. [Koller & Sahami, 1997]

Commodity level, but only up to *Class* or *Family* levels. Therefore, we build a hierarchical classifier, which actually consists of four classifiers, each of which is working on a correspondent level (see also [Chakrabarti et al., 1997], [Koller & Sahami, 1997], [Dumais & Chen, 2000], and [Agrawal & Srikant, 2001]).

3.4. Implementation

GoldenBullet is designed to provide widest access to product description classification service. So, we intended to make some kind of web service out of our tool. Our current version of the prototype is oriented on an “html like” user interface. Our main goal, concerning user interface design, was to provide fully functional and convenient for a user interaction environment and, at the same time, not to put too many requirements on the user–side software. Currently all what a user needs to use **GoldenBullet** prototype is an html browser that supports JavaScript 1.2.

The web service is provided by means of a client-server approach. So, the core of our tool is a server-side set of Java packages that implements the functionality and generates all interaction pages. The server side module was implemented as a web application. We use Java Servlets technology and its extension Java Server Pages to generate all client side user interface pages. All intelligent content of our prototype was implemented by means of Java 1.3. Due to the separation of the training-classification process in an “off-line” training step (computationally highly expensive) and an “on-line” classification step we managed to achieve acceptable performance.

4. Evaluation

The evaluation we report is based on around 40,000 real product descriptions that were already classified manually. They come from various vendors and cover a large number of categories in UNSPSC.[6] During the following we compare the performance of a number of algorithmic variants of our classifier.

4.1. Test Data Set

The data provide 41913 manually classified data described in France. Table 1. summarizes the population of the UNSPSC categories in the data set according to the UNSPSCv7.2.

Table 1. Population of UNSPSC

Category	Number of UNSPSC categories	5%	Number of instances in the most frequent categories
Segments	28	1	15233
Families	56	3	27407
Classes	150	7	24800
Commodities	421	21	28247

The table can be understood as follows: The number of UNSPSC categories reports the number of Commodities, Classes, Families, and Segments of UNSPSCv7.2 that are populated by some instances found in the data set. Second we analyzed the distribution of the data. We took around 5% of the covered Commodities, Classes, Families, and Segments and selected the ones with the highest coverage. It shows that 28247 data of the data set are in 21 commodities, i.e., around 60% of the data are in 21 of the around 15000 commodities. Similar only in one Segment are around 25% of all data. In total, we worked with 41913 product descriptions.

4.2. Accuracy Results for a Naïve Classifier

Up to a large number of the product descriptions in this test data set are represented by the name of the product models, such as “proliant”, “pentium”, “presario”, “carepaq”, “RA3000”, but do not use any of the functional terms of the product itself. In this case, our Naïve Classifiers are not capable to secure high accuracy. In fact, the accuracy is extremely low and reaches maximally 0,2%. Clearly such a classifier is without any use and we will describe in the following how training could make a different story.

4.3. Accuracy Results of the trained Algorithms

For training the algorithms we have chosen the following approach. A 60% random sample from product descriptions data set was used as training set, and the rest 40% data – as test set. We repeated the test based on several random splits of the data set. The results are reported in Table 2. We applied two quality measurements:

- The *total accuracy* asks whether the commodity recommendation of **GoldenBullet** with the highest

[6] The data were collected based on a cooperation with Hubwoo which is a MRO market place in France.

rating based on the product description matches the actual commodity of a product.

- The *"First 10 Accuracy"* asks whether one of the ten commodity recommendations of **GoldenBullet** with highest ratings based on the product description matches the actual commodity of a product.

In addition, we distinguished two modes for all algorithms. Either we treated the pre-classified product data or the enriched class descriptions (based on the pre-classified data) as documents that should be retrieved. In general, the bayesian classifier outperforms all other classifiers significantly.[7] Working directly with the pre-classified data works best for all algorithms. Only in regard to the *"First 10 Accuracy"* there is no difference for the bayesian classifier in this respect. In general, an accuracy between 78% to 88% looks rather convincing and easily outperforms and qualify equal with the quality of human classification.[8]

Table 2. Accuracy of trained (non-hierarchical) algorithms

Algorithm	Total Accuracy	First 10 Accuracy
VSM_I	60%	78%
VSM_C	28%	69%
KNN_I	45%	84%
KNN_C	29%	56%
NB_I	**78%**	**88%**
NB_C	59%	**88%**

We repeated the experiments for hierarchical versions of the algorithms, i.e., first a classification decision is taken at the segment level and then recursively on the families, classes, and commodity level. Against our initial intuition this lead to significant lowering of the accuracy of the classification algorithms. Obviously, too many wrong decision in the early steps of the classification process happens. The results are presented in Table 3.

Table 3. Accuracy of trained (hierarchical) algorithms

Algorithm	Total Accuracy	First 10 Accuracy
VSM_I	22%	41%
VSM_C	14%	27%
KNN_I	25%	22%
KNN_C	13%	15%
NB_I	38%	42%
NB_C	29%	42%

4.4. Summary

Untrained algorithms fail completely to provide any support in semi-automatic or automatic product classification. Trained algorithms can provide significant support. Up to 90% accuracy can be achieved based on learning from pre-classified example. As always the significant assumption of such an approach is the availability of representative pre-classified examples.

5. Conclusions and Future Works

Market places for B2B electronic commerce have a large economic potential. They provide openness and flexibility in establishing commercial relationships for their clients. In order to provide this service they have to tackle with serious obstacles. The most prominent one is concerned with integrating various styles to describe the content and the structure of the exchanged information. Product catalogues corresponds to large and complex domain ontologies and in the case of horizontal standards to upper-layer ontologies. Large modeling and mapping effort is required to handle these descriptions. Content manager have to structure, classify, re-classify, and personalize large volumes of data to make product descriptions automatically accessible via B2B market places.

GoldenBullet aims on mechanizing the classification process of product data. Accuracy rates between 70% and 98% indicate that this process can be mechanized to a degree where severe cost reduction can be achieved which is a pre-requisite for scalable E-commerce. The success of **GoldenBullet** is based on the combination of natural language processing, information retrieval, machine learning and the use large volumes of manually classified data. Future versions of the **GoldenBullet** will provide more advanced features as explained in the

[7] Compare similar results of [Agrawal & Srikant, 2001].

[8] That is, higher accuracy would just mean over-fitting to human classification decisions that also have a significant error rate which we encountered in labour intensive manual checks.

following.

Recently, **Support Vector Machines** have been shown to be a very useful tool for text categorization (cf. [Burges, 1998], [Cortes & Vapnik, 1995], and [Joachims, 1998], and [Dumais & Chen, 2000]). We want to explore the applicability of this algorithmic paradigm for our specific product classification task.

Multi-linguality (i.e. the product catalog and the product classification standard are described in different languages) is a severe requirement for E-commerce in Europe. Currently, **GoldenBullet** supports English and French. An extension to further languages is a pre-requisite for open and flexible E-commerce.

Multi-standard classification is an important issue for open and flexible E-commerce (cf. [Agrawal & Srikant, 2001], [Madhaven et al., 2001], and [Schulten et al., 2001]). Currently, market places establish one "standard" to make data of their clients accessible. However, openness and flexibility of internet-enabled trading demands multi-standard classification. Besides UNSPSC there are other "standards" like UCEC[9], ecl@ss[10], and RosettaNet[11] that are widely used in certain domain-specific or geographical areas of E-commerce. Providing mechanized support in reclassifying product data in an additional classification schema is a demanding feature for realizing the full potential of E-commerce.

GoldenBullet applies successfully information retrieval and machine learning techniques to the problem of automatically classifying product description. **GoldenBullet** will also challenge other existing severe problems for B2B market places, such as mapping and reclassifying product descriptions according to different product code systems[12] and to personalize views on product data for divergent customers.

References

[Agrawal & Srikant, 2001]
R. Agrawal and R. Srikant: On Integrating Catalogs. *In Proceedings of the Tenth International World Wide Web Conference (WWW2001)*, Hong Kong, May 2001.

[Burges, 1998]
C. J. C. Burges: A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2(2):955-974, 1998.

[Chakrabarti et al., 1997]
S. Chakrabarti et al.: Using taxonomy, discriminants, and signatures for navigating in text databases. In *Proceedings of the 23rd VLDB Conference*, Greece, 1997.

[Cheeseman and Stutz, 1995]
P. Cheeseman and J. Stutz: Bayesian classification (AutoClass): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), *Advances in knowledge discovery and data mining*, The AAAI press, Menlo Park, 1995.

[Cortes & Vapnik, 1995]
C. Cortes and V. Vapnik: Support-Vector Networks, *Machine Learning*, 20:273-297, 1995.

[Dumais & Chen, 2000]
S. Dumais and H. Chen: Hierarchical Classification of Web Content. In *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval,* pp. 256-263, Athens, Greece, August 2000.

[EDIFACT, 1999]
United Nation: *UN/EDIFACT-Directory.* http://www.unece.org/trade/untdid, 1999.

[Fensel, 2001]
D. Fensel: *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*, Springer-Verlag, Berlin, 2001.

[Fensel et al., 2001]
D. Fensel, Y. Ding, E. Schulten, B. Omelayenko, G. Botquin, M. Brown, and A. Flett: Product Data Integration in B2B E-commerce, *IEEE Intelligent System*, 16(3), 2001.

[Gomez-Hidalgo & Rodriguez, 1997]
J. M. Gomez-Hidalgo and M. B. Rodriguez: Integrating a lexical database and a training collection for text categorization. In the *Proceedings of ACL/EACL (the Association for Computational Linguistics/European Association for Computational Linguistics: Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications)*, Madrid, Spain, July, 1997.

[Joachims, 1998]
T. Joachims: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In C. Nedellec and C. Rouveirol (eds.), *Machine Learning: ECML-98, Proceedings of the 10th European Conference on Machine Learning,* Lecture Notes in Artificial Intelligence, LNAI 1398, pages 137-142, 1998.

[Koller & Sahami, 1997]
D. Koller and M. Sahami: Hierarchically classifying documents using very few words. In *Proceedings of the International Conference on Machine Learning*, vol 14, Morgan-Kaufmann, July 1997.

[Madhaven et al., 2001]

[9] http://www.ucec.org

[10] http://www.eclass.de

[11] http://www.rossettanet.org

[12] See [Agrawal & Srikant, 2001] and [Navathe et al., 2001].

J. Madhaven, P. A. Bernstein, and E. Rahm: Generic Schema Matching with Cupid. In *Proceedings of the 27th VLDB Conference*, Roma, Italy, 2001.

[Mitchell, 1997]
T. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[Navathe et al., 2001]
S. Navathe, H. Thomas, M. Satitsamitpong, and A. Datta: A Model to Support E-Catalog Integration. In *Proceedings of the 9th IFIP 2.6 Working Conference on Database (DS-9), Semantic Issues in e-commerce Systems*, Hong Kong, April 2001.

[Ribiero-Neto & Baeza-Yates, 1999]
B. Ribiero-Neto and R. Baeza-Yates, *Modern Information Retrieval*, Addison Wesley, 1999.

[Salton et al., 1975]
G. Salton, A. Wong, and C. S. Yang: (1975): A vector space model for automatic indexing, *Communications of the ACM*, 18(7):613-620, 1975.

[Schulten et al., 2001]
E. Schulten, H. Akkermans, G. Botquin, M. Dörr, N. Guarino, N. Lopes, and N. Sadeh: The ecommerce product classification challenge, *IEEE Intelligent systems*, 16(4), 2001.