GoldenBullet in a Nutshell

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

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. 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 sketch 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.

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-dated 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 discusses future directions.

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. One of the major challenges 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].

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. A typical content management solution provider 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. 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[1] (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. Therefore this process is costly, however, a high quality is important to ensure maintainability and visibility of product information.

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)*[2]. It takes UNSPSC as a starting point and refines it by attributes. Rosetta Net[3] 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. Because different customers will make use of different classification schemas the product information must be classified and described according to several schemas.

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.

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 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 15,000 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 so as to improve the classification accuracy.

In the following we will first introduce UNSPSC which is the product standard we use in automated product classification. Then we discuss the overall functionality of **GoldenBullet**. This is followed by a more detailed discussion of the core component of **GoldenBullet**, i.e., various classifier variants. We conclude with some additional strategies that help to improve the overall performance and add some more details on the implementation of it.

The Universal Standard Products and Services Classification (UNSPSC) is an open global coding system that classifies products and services. It is 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 15,000 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.

Cataloguing product description according to UNSPSC is a big burden carried by B2B market place vendors. Currently it is mainly done manually. Achieving semi-automatic or

1. http://www.un-spsc.net and http://www.unspsc.org
2. http://www.ucec.org
3. http://www.rosettanet.org/

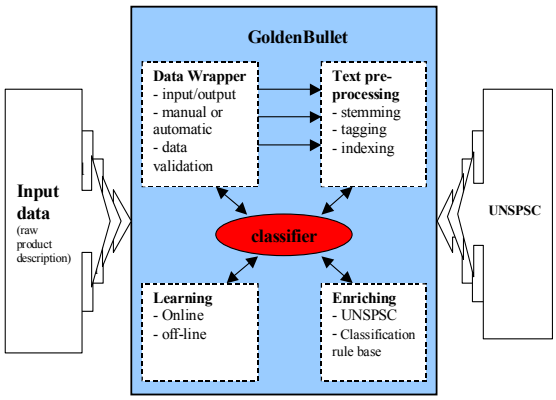

Figure 1. Overview on **GoldenBullet**.

automatic support in cataloguing product description is a significant breakthrough. **GoldenBullet** is a prototype for deploying information retrieval and machine learning methods to classify product description semi-automatically or automatically: Data input and export facilities; text processing techniques; classification of product data; and learning and enrichment of product classification information (see Figure 1).

A **wrapper factory** gathers various wrappers to convert raw data description from external formats (Database, Excel, XML-like, formatted plain text,...) to internal format, and final results to preferable output format or 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.

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). A stop word list has been generated, updated and extended during the whole process. Currently, **GoldenBullet** can handle English and French product data.

Figure 2 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 and 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:

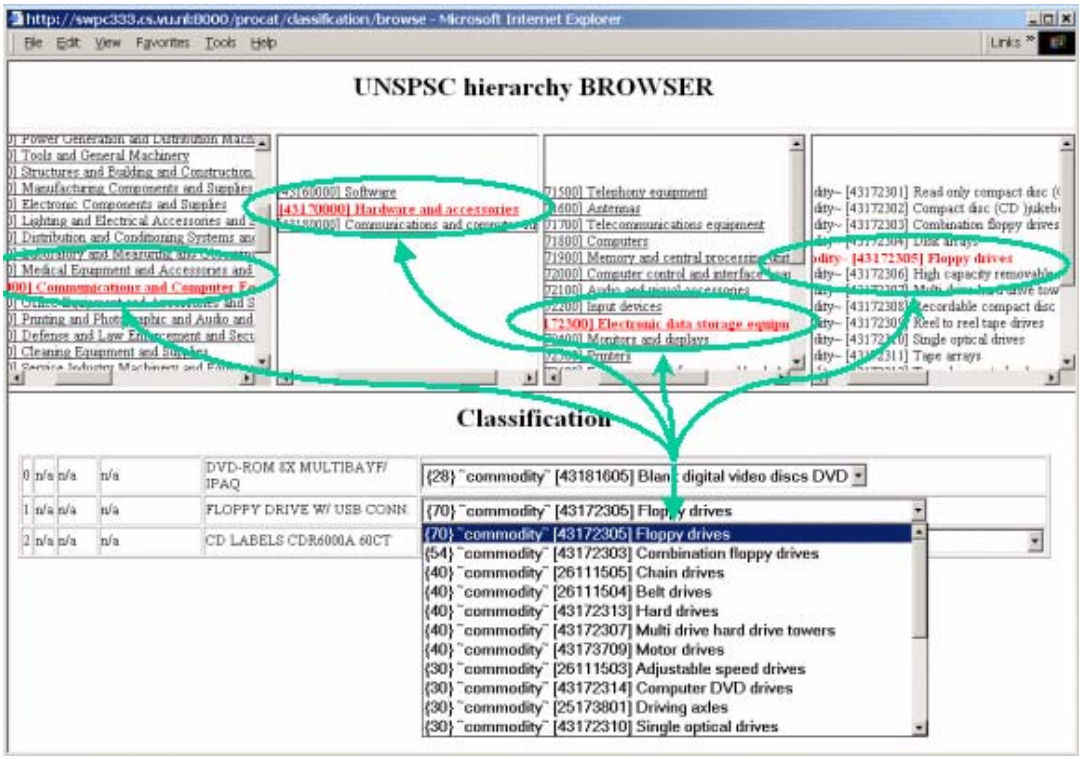

Figure 2. A screenshot of **GoldenBullet**.

- 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 corresponding 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. Therefore, we employed various strategies to achieve a more reasonable and workable result. Basically we employed different retrieval strategies and we made use of large volumes of manually classified data to improve the performance.

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.

Another we implemented is based on the k-Nearest Neighbor method **KNN**.[4] 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. We can distinguish two modes in regard to whether the algorithm works directly on the pre-classified product data or on enriched class descriptions.

The final paradigm we employed is the machine learning paradigm. 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 [Cheeseman and Stutz, 1995] 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. The Naïve-Bayes **NB** is a standard text classification algorithm, with a long successful application history.

4. http://citeseer.nj.nec.com/cs?q=knn&submit=Search+Citations&cs=1

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. Therefore, we build a hierarchical classifier, which actually consists of four classifiers, each of which is working on a correspondent level.

GoldenBullet is designed to provide widest access to product description classification service. Our current version of the prototype is oriented on an "html like" user interface. Currently all what a user needs to use **GoldenBullet** prototype is any html browser that supports Java Script 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.

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.[5] During the following we compare the performance of a number of algorithmic variants of our classifier.

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 (between 0.1 and 0.2%). Clearly such a classifier is without any use and we will describe in the following how training could make a different story.

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 1. We applied two quality measurements:

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

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

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. 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.[6]

Table 1. 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 decisions in the early steps of the classification process happen.

Conclusions

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. **GoldenBullet** aims on mechanizing the classification process of product data. Accuracy rates between 70% and 90% 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 such as multi-linguality and multi-standardization.

- **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.
- **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 and to personalize views on product data for divergent customers (cf. [Schulten et al., 2001], [Corcho Garcia & Gomez-Perez, 2001]).

References

[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.

[Corcho Garcia & Gomez-Perez, 2001] O. Corcho Garcia and A. Gomez-Perez: Solving Integration Problems of e-commerce Standards and initiatives through ontological mappings. In *Proceedings of the IJCAI'01 Workshop on Ontologies and Information Sharing*, Seattle, Washington, USA, August 2001.

[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)*, Madrid, Spain, July, 1997.

[Ribiero-Neto and 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.

6. 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.