Forecasting Consumer Interest in New Services using Semantic-aware Prediction Model: the Case of YouTube Clip Popularity

Luka Vrdoljak[1], Vedran Podobnik[2], Gordan Jezic[2]

[1]Erste & Steiermärkische Bank, Croatia – lvrdoljak@erstebank.com
[2]University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia – {vedran.podobnik, gordan.jezic}@fer.hr

Abstract. With intense increase in number of competing service providers in the information and communication sector, companies must implement mechanisms for forecasting consumer interest in new services. Common growth models provide the mechanisms for modelling and predicting acceptance of a certain service. However, they have two shortcomings: i) limited precision; and ii) a short, but yet existing, time delay. By using semantic reasoning for detecting similarities between services already on a market and ones that are just to be introduced, it is possible both to increase forecasting precision and eliminate the time delay caused by the need to collect a certain amount of data about the new service before a prediction can be made. The proposed semantic-aware prediction model is elaborated on a case of forecasting YouTube clip popularity.

Keywords: Consumer Relationship Management, Consumer Managed Relationship, Forecasting, Growth Models, Semantic Reasoning, YouTube

1 Introduction

Considering the increasing number of competing service providers and services on the information and communication market, companies must focus on maintaining consumer satisfaction. In order to do so, service providers must observe their consumers individually, rather than seeing them just as a part of a certain market niche. Such individual and personalized approach can be recognized on the market through two most common concepts: *Consumer Relationship Management* (CRM) and *Consumer Managed Relationship* (CMR) [1][2]. These two concepts reside on three basic ideas: i) consumer experience management (CEM), ii) real-time analysis, and iii) technology used for cost decreasing and creating a consumer-oriented environment [3][4][5].

Most companies react only when a number of consumers decreases (i.e. churn rate dominates over growth rate). However, by then it is usually too late to intervene. On the other hand, using *Predictive Analysis* (PA) can lead to a proactive consumer retention strategy [4]. By analysing consumer habits, expenditure and other behaviour

Vrdoljak, Luka; Podobnik, Vedran; Jezic, Gordan.
Forecasting Consumer Interest in New Services using Semantic-aware Prediction Model: the Case of YouTube Clip Popularity.
Lecture Notes in Computer Science. 7327 (2012); 454-463.

patterns, forecasting models can determine the probability of a decrease in consumers' interest in a certain service, or even the potential interest in a service that has yet to be introduced in the market [6]. Additionally, predictions could be improved with the use of Semantic Web technologies which enable detection of similar services already on the market.

In this paper, we will first make an insight in the common forecasting models and state-of-the-art technologies for semantic service profiling (Section 2). Then we will present our proposed system through key processes (Section 3) and its architecture (Section 4). In Section 5 we will conduct an evaluation of our ideas on YouTube clip popularity forecast. Section 6 concludes the paper and presents the future research planned to result in original scientific contribution of a doctoral candidate.

2 Related Work

After a long era of easily predictable fixed voice telephone services, information and communication industry has come to a period of intensive introduction of a very wide spectrum of numerous new services [8]. Rapid technological development and liberalisation have made the information and communication market a very dynamic environment where forecasting is becoming increasingly important. By understanding data patterns during information and communication services' life-cycles, a service provider can perform optimal business planning of its capacities, investments, resources (e.g. human potentials and equipment), marketing and sales. However, there is always the problem of bridging the gap between collected historical data and the anticipated value in the future due to the lack of reliable input data for forecasting.

2.1 Service Growth Models

Every service's life-cycle (SLC) consists of phases shown in Fig. 1 [9]: *development*, *introduction*, *growth*, *maturity* and *decline*. A typical service during its life-cycle passes through specific phases of market adoption, which can be observed through the number of service users. Understanding these phases in a SLC is especially important for highly competitive market environments and particularly for services based on emerging technologies. Numerous researches have resulted in a conclusion that these phases can be described by mathematical models which will be briefly explained here.

Growth models mathematically describe patterns of growth in nature and economy, illustrate how a certain environment reflects on the growth, as well as enable future growth forecasting. Particularly, diffusion of new ideas and technology, market adoption of new products and services, as well as allocations of restricted resources has characteristic S-shaped (sigmoidal) growth. Two most commonly used S-shaped models for initial phases of a SLC (i.e. development, introduction, growth) are the *Logistic* and the *Bass* models. Later phases of a SLC require more complex models (e.g. Bi-Logistic growth model) [8]. This paper will be focused on the services in their initial SLC phases.

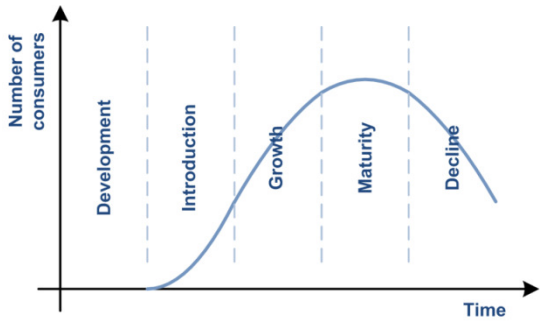

Fig. 1 - Information and communication service life-cycle

Logistic Model

The logistic model $L(t)$ is best used for describing growth of the number of service consumers in time in a closed market, isolated from other services. The model is defined with three parameters: M – market capacity, a – growth rate, and b – time shift parameter, as is shown in (1) [10].

$$L(t; M, a, b) = \frac{M}{1 + e^{-a(t-b)}} \qquad (1)$$

The logistic model is a widely used growth model with numerous useful properties for technological and market development forecasting. During the first phase, growth of the logistic model is exponential, but later negative feedback slows the gradient of growth as the number of consumers approaches the market capacity limit M. Individual impact of each parameter can be seen in Fig. 2.

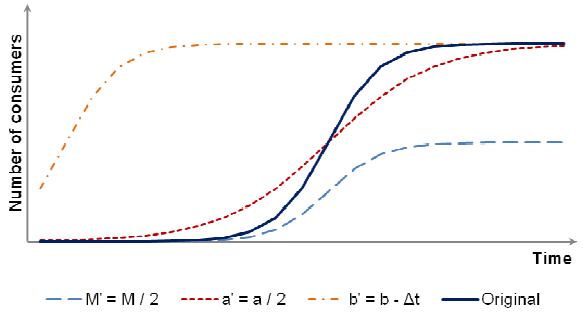

Fig. 2 - Interest in a service described by the *Logistic model* with different parameters

Bass Model

The most common model for describing new service diffusion is the Bass model [11]. The Bass model $B(t)$ corrected the deficiency of simple logistic growth (slow growth and no point where $L(t)$ equals zero) by taking into account the effect of innovators via coefficient of innovation p. The model divides a population of M adopters in two categories: *innovators* (with a constant propensity to purchase) and *imitators* (whose

propensity to purchase is influenced by the amount of previous purchasing). Bass diffusion model is defined by the following four parameters:

- M – market capacity;
- p – coefficient of innovation, $p > 0$;
- q – coefficient of imitation, $q \geq 0$, and
- t_s – the moment of service introduction, $B(t_s) = 0$.

These parameters define the model as shown in (2). The Bass model has a shape of S-curve, as does the Logistic model, but the curve is shifted down on the y-axis. Fig. 3 shows the effects of different values of parameters p and q on form of S-curve, with fixed values for M and t_s [8].

$$B(t; M, p, q, t_s) = M \frac{1 - e^{-(p+q)(t-t_s)}}{1 + \frac{q}{p} e^{-(p+q)(t-t_s)}} \qquad (2)$$

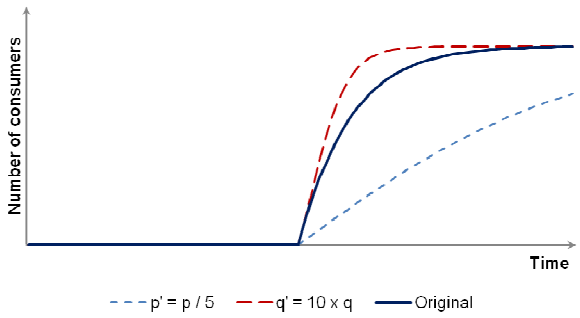

Fig. 3 - Interest in a service described by the *Bass model* with different parameters

The Bass model is widely used for the long-term forecasting of a new service market adoption when interaction with other services can be neglected. Furthermore, the Bass model can be used when limited or no data is available (i.e. market adoption forecasting prior to service launch), which is also a focus of our paper. However, while best practices in such cases recommend obtaining model parameters via subjective judgmental assumptions and placing them in optimistic-pessimistic intervals we propose a different, more objective approach. Namely, researches have shown that services with similar characteristics tend to have similar parameters when their growth is represented with the Bass (or Logistic) model [12]. Taking this into account we propose using Semantic Web technologies to profile services and thus enable computers to autonomously calculate the level of similarity between any pair of services.

2.2 Semantic Service Profiling

Semantic markup has proven to be a very efficient method for resource matchmaking and ultimately better recognition of consumer needs. The Semantic Web is a concept in

which knowledge is organized into conceptual spaces according to meaning, and keyword-based searches are replaced by semantic query answering. Semantic Web languages, such as Resource Description Framework (RDF), RDF Schema (RDFS) and the Web Ontology Language (OWL), are used to provide structure for resource description (i.e. service profiles) [7].

Using various query languages, based on Structured Query Language (SQL) syntax, it is possible to perform semantic data extraction, thus enabling efficient matchmaking of service profiles once they have been created according to a certain standard. Such matchmaking enables service comparison to be performed according to true, semantic similarities, rather than keyword matchmaking (i.e. YouTube clip tags).

3 Service Modelling System

In order to perform new service growth forecasting it is necessary to ensure mechanisms for four basic processes: i) creating semantic profiles based on service description, ii) mapping existing services' historical data into growth models (i.e. Bass or Logistic model), iii) comparing newly introduced service with existing services, and finally, iv) calculating newly introduced service growth model and corresponding parameters.

Semantic Profiling

First, it is necessary to create an ontology that relates information and communication service descriptions (e.g. tags, multimedia clip reproduction quality and length, etc.) to semantic profiles that can be autonomously processed by computers [7]. It is not possible to create comprehensive service ontology because the number and variety of services on the market increase on an hourly basis. Therefore, it is necessary to ensure a simple ontology upgrade process to enable description of new services, as well matchmaking of new services with those already in the market.

Modelling Existing Services

Number of consumers of a certain services, when observed through time, forms a rather irregular set of discrete data points. Our system must transform this stochastic set of data into a smooth S-curve that approximates the actual numbers with a satisfactory degree of deviation (Fig. 4). Finding the correlation is performed in two steps: recognizing the correct model (e.g. Bass or Logistic model) and calculating the corresponding parameters. For a data model defined by k parameters it is necessary to have at least k known data points. When there is exactly k known data points, the parameters are the solutions of a system of equations. System is usually nonlinear, so iterative numerical methods need to be performed for its solution. When there are more than k data points known, the weighted least squares method is the most commonly used. The objective is to minimise sum of squared differences between actual data points from a real world and evaluated points from a model [8].

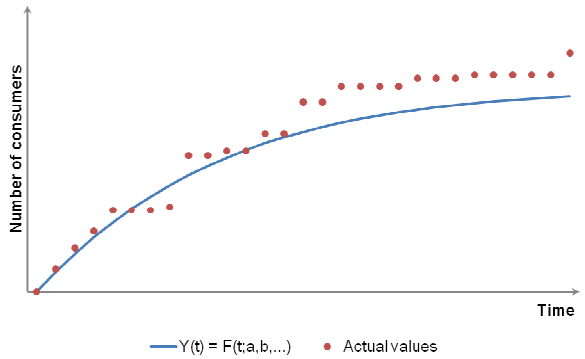

Fig. 4 - Transforming actual data points into a model defined as Y(t)

Introducing and Modelling Newly Introduced Services

The final goal of our system is to forecast consumer interest in newly introduced services by calculating its growth model. In order to achieve that it is firstly necessary to see where the newly introduced service fits in the existing set of services on the market. We propose using semantic matchmaking to detect most similar services [7]. Once we have identified similar services it is possible to choose the most appropriate growth model (i.e., the most common growth model among similar services) and calculate the chosen model parameters based on parameter values of the similar services, taking corresponding semantic similarities as weight factors (i.e. the most similar services are the most important while calculating parameter values). This is shown in Fig. 5.

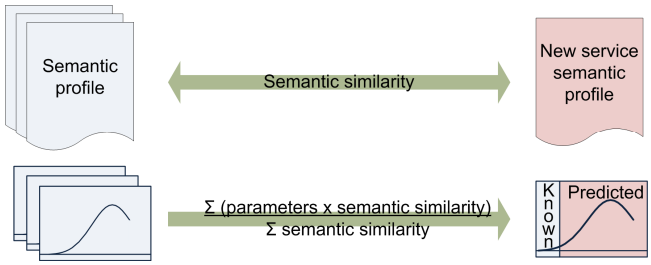

Fig. 5 - The process of predicting consumer interest in newly introduced service

4 System Architecture

The system we propose in this paper consists of three main entities: i) *service provider*, ii) *semantic repository*, and iii) *agent-based application* (Fig. 6). The service provider (e.g. YouTube) offers its consumers information and communication services (e.g. multimedia clips) via Internet. Each service is characterised with its description (e.g. identification, category, author, resolution, tags) and data that defines

the number of its consumers (e.g. multimedia clip viewers) from the moment it was introduced on the market.

The `Semantic Profiling Agent` transforms service descriptions into semantic profiles. The `Service Modelling Agent` recognizes the adequate growth model for each existing service with sufficient data and calculates the parameters accordingly. The semantic profile and growth model information (for each service) are then stored into a semantic repository so they can later be used for semantic matchmaking with newly introduced services, as well as calculating growth model of newly introduced services. The `New Service Modelling Agent` uses the information stored within the semantic database and calculates the model when a new service is introduced and there is none or insufficient data to calculate the model from its data points.

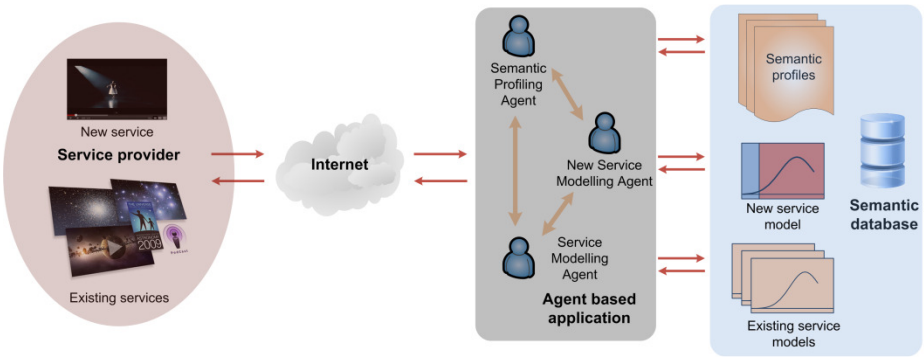

Fig. 6 - Architecture of an agent-based system for predicting consumer interest in newly introduced services

5 Proof-of-Concept Scenario: Calculating a Growth Model for a Newly Introduced YouTube Clip

In this section we will present the functionalities of our proposed system. We will use YouTube as a proof-of-concept service provider and multimedia clips as a proof-of-concept information and communication services. YouTube clips are divided into categories (i.e. music, film & animation, education, etc.). Also, each clip has a short description and a set of tags suggested by the author. This information is translated into a semantic profile as described in Section 4. An example of a complete service profile is shown in Fig. **7**. The profile consists of four main parts: *identification* (i.e. name, author, URL, etc.), *technical characteristics* (i.e. video clip resolution, frames per second, duration, etc.), *keywords* retrieved from tags (i.e. actor names, director name, content keywords, etc.), and calculated *growth model information* (i.e. model, parameters).

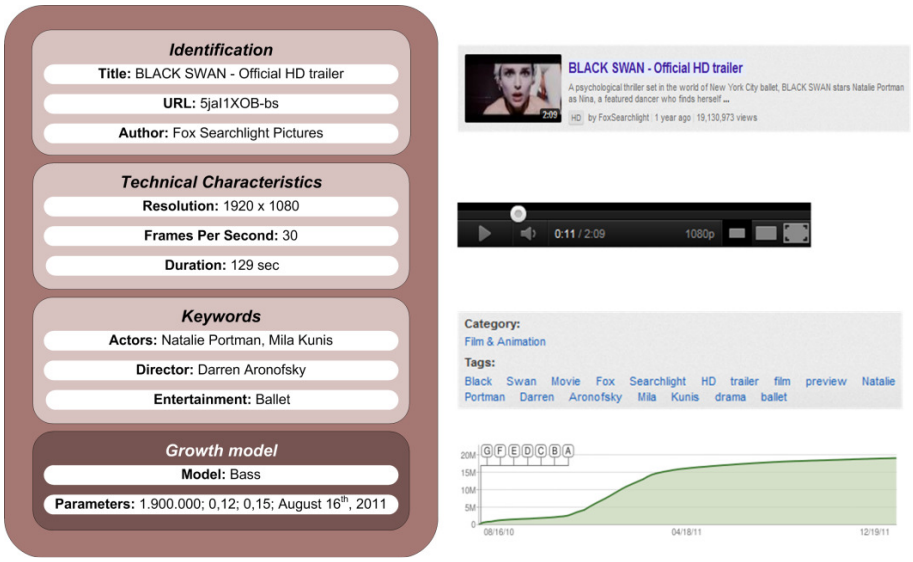

Fig. 7 - Semantic profiling of a YouTube clip[1]

Service Matchmaking

Comparison of YouTube clips is based on semantic matchmaking that is performed by the `New Service Modelling Agent` [7]. Each semantic profile consists of attributes and joint values. Semantic matchmaking is based on numeric evaluation of categories clips belong to, basic data type attributes (i.e. resolution, duration), and class instance attributes (i.e. actors, director). Such numeric evaluation results in a final similarity level between 0 and 1.

Semantic matchmaking is first performed between clips' categories. For example, two YouTube clips from *Sports* category are more likely to have similar parameter values than two clips from different categories. After that each common attribute is compared. Numeric value comparison is the quotient of the smaller and the larger number. For example, if we compare clips with durations of 150 and 300 seconds, their similarity is 0.5. String and binary value comparison result is 1 if the values are identical and 0 otherwise. Class instance attributes are compared observing classes' positions in the ontology class hierarchy, where the result is 1 divided by the number of steps between them [7]. For example, if we compare two clips where one has *Ballet* in the *Keyword* section, and the other has *Opera* in the same section, with both *Opera* and *Ballet* representing instances of class *Entertainment*, their similarity will be 0.5.

Once all the attributes have been compared, final profile similarity is calculated using *weighted arithmetic mean* method where weight factors are custom values that represent the importance of each attribute. For example, clip category should have a much higher weight than *Frames per Section* attribute. More detailed insight in semantic matchmaking is presented in [7] and [13].

[1] The information was retrieved from: *http://www.youtube.com/watch?v=5jaI1XOB-bs* (accessed: December 19th, 2011)

Growth Forecasting for Newly Introduced Services

If we were to look back at the time when the *Black Swan Trailer* had just been uploaded, a whole different modelling process would have occurred. When a new video clip is introduced it is compared to all existing clips in the system by the `New Service Modelling Agent`. After semantic matchmaking results are obtained we choose a certain number (e.g. five) of clips most similar to the newly introduced clip. An example of calculated similarities and Bass model parameters for most similar clips is shown in Tab. 1.

Tab. 1 - Sample clip data

	Semantic similarity	M	p	q
YouTube Clip 1	0.754	2 050 200	0.125	0.132
YouTube Clip 2	0.726	2 433 150	0.139	0.078
YouTube Clip 3	0.719	1 510 220	0.105	0.112
YouTube Clip 4	0.698	2 601 500	0.098	0.102
YouTube Clip 5	0.695	1 896 000	0.117	0.087
YouTube Clip 6	0.631	2 100 500	0.058	0.051
YouTube Clip 7	0.624	4 300 400	0.133	0.035
YouTube Clip 8	0.523	1 500 000	0.081	0.125
YouTube Clip 9	0.496	500 000	0.048	0.193
YouTube Clip 10	0.478	215 000	0.052	0.032
...	...	...	...	...

Parameters for the newly introduced clip are calculated as follows (according to the rule from the Fig. 5):

$$M = \frac{\sum_{i=1}^{5}(s_i \times M_i)}{\sum_{i=1}^{5} s_i} = 2\,096\,808 \qquad (3)$$

$$p = \frac{\sum_{i=1}^{5}(s_i \times p_i)}{\sum_{i=1}^{5} s_i} = 0.117 \qquad (4)$$

$$q = \frac{\sum_{i=1}^{5}(s_i \times q_i)}{\sum_{i=1}^{5} s_i} = 0.103 \qquad (5)$$

where s_i represents the semantic similarity between the new clip and YouTube clip i, while M_i, p_i and q_i are Bass model parameters of clip i. Once parameters M, p and q are calculated it is possible to approximate the number of viewers the new clip should reach in near future. As was mentioned earlier, basic Bass model works best if used on initial growth modelling. Latter stages of SLC require more complex models.

6 Conclusion and Future Work

In this paper, we propose a multi-agent system for forecasting consumer interest in new services. The innovativeness of our proposal can be recognized in using semantic reasoning for enhancing newly introduced service growth modelling. The semantic reasoning is particularly helpful when insufficient data about new service popularity is available – semantic reasoning enables us to substitute missing data for parameter

calculation with the data from similar services already on the market. Such approach should enable service provider to perform pre-market forecasting in order to determine whether the service has its place in the market or it is destined for failure.

Our future research will be focused on the following three challenges. The first challenge is improving the implemented semantic reasoning mechanism. Key tasks which correspond to this challenge are improving scalability of semantic matchmaking algorithm and creating mechanisms for automated service profiling (e.g. automated transformation of YouTube clip tags into semantic profiles). The second challenge is implementing a more generalized model that will be applicable over complete service life-cycle (not just for the initial life-cycle phases). The final challenge is verification of our proposed system on various information and communication services (e.g. forecasting consumer interest in news based on news diffusion through a social network).

References

1. Vrdoljak, L., Bojic, I., Podobnik, V., Jezic, G., Kusek, M.: Group-oriented Services: A Shift toward Consumer-Managed Relationship in Telecom Industry. Transactions on Computational Collective Intelligence. 2 (2010); 70-89.
2. Sreedhar, D., Manthan, J., Ajay, P., Virendra, S.L., Udupa, N. Customer Relationship Management and Customer Managed Relationship - Need of the hour. http://www.pharmainfo.net/ (April 2011)
3. Schmitt B. Customer experience management: a revolutionary approach to connecting with your customers. John Wiley and Sons, Hoboken, New Jersey, USA, 2003.
4. Girish, P.B. How Banks Use Customer Data to See the Future. http://www.customerthink.com/ (April 2011)
5. Shin, N. Strategies for Generating E-Business Returns on Investment. Idea Group Inc, 2005.
6. Rygielski, C., Wang, J.C., Yen, D.C. Data mining techniques for customer relationship management. Technology in Society. 24 (2002); 483-502.
7. Vrdoljak, L. Agent System based on Semantic Reasoning for Creating Social Networks of Telecommunication Service Users. Diploma thesis, University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, 2009.
8. Sokele, M. Analytical Method for Forecasting of Telecommunications Service Life-Cycle Quantitative Factors. Doctoral Thesis. University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, 2009.
9. Sokele, M., Hudek, V. Extensions of logistic growth model for the forecasting of product life cycle segments. Advances in Doctoral Research in Management (ed. L. Moutinho), World Scientific Publishing. 1 (2006); 77-106.
10. Sokele, M., Growth models for the forecasting of new product market adoption. Telektronikk. 3/4 (2008); 144-154.
11. Bass, F. A new product growth for model consumer durables. Management Science. 15 (5) (1969); 215-227.
12. Tellabs. Forecasting the Take-up of Mobile Broadband Services. White Paper. 2010.
13. Li, Y., McLean, D., Bandar, Z., O'Shea, J., Crockett, K. Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering. 18 (8), (2006); 1138-1150.