

Dynamic Population Segmentation in Online Market Monitoring

Norbert Walchhofer * Karl Anton Froeschl †

Milan Hronsky * Kurt Hornik ‡

Preprint of a contribution sent to

Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V., Dresden

20. October 2009

Abstract

The objective of the SEMAMO (Semantic Market Monitoring) project is to make use of the increasingly growing information available at Web-based sales and marketing channels for market research, using semi-automatic analysis driven by application domain models. The assumptions are that (i) the Web may serve as a representative "picture" of the reality, and (ii) the respective online channels map salient market developments and (iii) all of this accurately and in a timely manner.

Limited server requests and market specific access structures of Web portals inhibit both full scans of sampling populations and random selection of sampled offers. Further, product feature categories entail multiple classifications within offer *clusters* (e.g., geography in tourism). Therefore, SEMAMO proposes an *adaptive* sampling strategy dealing simultaneously with (i) the dynamics of the population frame, (ii) price dynamics, and (iii) multiple (fuzzy) classifications of offered products.

The paper discusses a heuristic method of dynamically segmenting monitored offer populations to stratify online data harvesting depending on both observed price changes and information relevance, and outlines the mechanics of harvest schedule derivation.

Keywords: population segmentation, adaptive sampling, information extraction.

*EC3—e-commerce competence center, Vorlaufstr. 5/6, 1010 Vienna, Austria

†University of Vienna, Institute of Scientific Computing, Universitaetsstr. 5, 1010 Vienna, Austria

‡Vienna University of Economics and Business, Augasse 2-6, 1090 Vienna

1 Introduction

The transparency of e-markets and increasing market dynamics call for more responsive and encompassing approaches towards the monitoring of markets and competition. A natural response to this overall development, advanced information technology [1]—and particularly semantic technologies—provide an unprecedented means to expand both the scope and speed of market observation by reducing the cost of information procurement and, thus, improving competitive decision making. In this respect, SEMAMO [2] arguably extends the range of current business intelligence methodologies and solutions by designing and implementing a (semi) automatic market monitoring framework, capitalizing on semantically enriched models of online information extraction.

The ensuing framework is applied to a leading e-commerce domain, tourism, providing (i) a fairly challenging test-bed in terms of market complexity [3], (ii) an information-rich environment comprising a multitude of online marketing channels, and (iii) structural peculiarities such as volume and access constraints restricting actual data retrieval. Hence, efficient monitoring of online markets depends on a dynamic allocation of access resources, adjusting the observation and analysis effort to varying market conditions.

Section 2 of this paper briefly relates the methodology employed in SEMAMO to preceding work in information retrieval, sampling in online contexts, and modeling of dynamic phenomena such as market prices. Next, Section 3 sketches the supposed model of price dynamics in online markets and introduces the heuristic evidence-based SEMAMO approach towards adaptive market segmentation reflecting the similarity of price change patterns. Referring to the fundamental feedback loop governing the adaptive SEMAMO data harvesting scheme, Section 4 then describes the proposed harvest balancing approach to derive (by combinatorial optimization) feasible data harvest schedules representing both the dynamics and economic utility of market information gathered iteratively. The concluding section presents a summary of the already achieved state of project development, and indicates further project challenges.

2 Related Work

Online market monitoring gathers product information (quality, price) across various online portals by extracting market information [4] from heterogeneous semi-structured sources [5], using specifically designed wrapper tools [6, 7]. Contrary to many attempts of document retrieval [8] seeking to optimize precision and recall for a wide range of conceivable queries, SEMAMO continually observes a set of deeply structured objects of interest over time.

It turns out empirically that, in many applications, (online) markets typically feature (discrete) jump processes rather than continuously varying prices. Thus, except for its purpose to monitor price dynamics of products within identified markets, the SEMAMO task resembles the tracking of occasional changes in (large) document sets [9]. Efficiency considerations suggest exploiting evidence of change dynamics for the sake of parsimonious observation; accordingly, SEMAMO capitalizes on an adaptive sampling model reflecting (expected) frequencies of price changes expressed in terms of Poisson-distributed latencies [10, 11]. However, the online habitat of SEMAMO inhibits a straight application of proven sampling methods [12]; notably, the populations to sample from are explored in a piecemeal fashion as inherent part of the information extraction process proper.

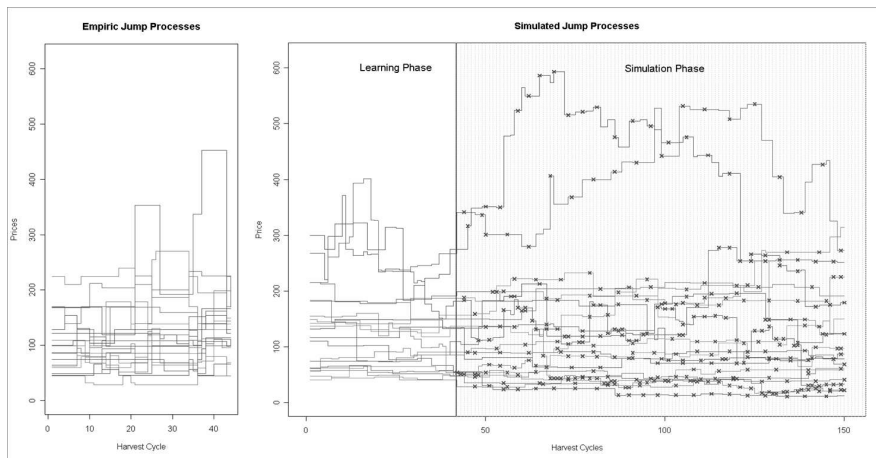


Figure 1: Empiric and simulated jump processes showing active harvest heuristic’s mode of operation by crosses marking update points

3 Sensor Binning Based on Price Dynamics

By definition, a market monitor aims at tracking price levels as well as their change dynamics of products on offer, for varying degrees of aggregation. Typically, such as in tourism, the offers within a single market are placed on different sales channels—Web portals, in particular—simultaneously; apparently, the same product may appear on multiple portals, with possibly differing prices. Thus, as an analytical unit of observation, it is reasonable to choose an individual product—such as a hotel room, or a package tour, to book—irrespective of multiple portal occurrences, implying that, over time, offer prices may vary reflecting changing market conditions. In what follows, the technical term “sensor” is used to denote a particular offer representing a triplet of (i) channel (Web portal), (ii) market aggregator, and (iii) the individual product on sale as such enabling statistical aggregation over, as well as comparison between, each of these sensor components. However, in an online context, target populations are rather ill-defined because of (i) the pace of change and, as a consequence of this, (ii) the difficulty of actually tracking all population members.

Methodologically, SEMAMO implements a directed data flow (typically, a set of online stores, or portals, in a given market) from—pre-selected—Web sources of raw online observation data towards aggregated business reports. Starting each harvest cycle with a data harvesting component, consisting of wrappers, and a data transformation unit [13] attached, cleansed data is rectified into a regularized representation of price series for the offers monitored. Aligned across data sources in terms of multi-dimensional data warehouse structures, regularized and accumulated price data are ready for a variety of statistical analyses according to customer-defined market reports. Additionally, parameters estimated from accumulated harvest data are used to *adaptively* drive the iterated harvesting of online data.

To illustrate typical price dynamics, the left-hand side of Figure 1 exhibits a sample of 20 price series taken from the SEMAMO test domain of hotel room offers, tracked over some 40+ harvest cycles. The right-hand part of Figure 1 contrasts these real price series with simulated ones, using Poisson-distributed jump processes (with $\lambda \sim \gamma(5, 5)$ chosen randomly for each price series, a magnitude random step size $\theta \sim \beta(0.4, 4)$, and the sign of step change also chosen randomly 50:50, in this sequence; starting prices have been generated exponentially distributed with $\lambda = 90$ within the interval $[40, 300]$).

3.1 Harvest Adaptation

Because of technical reasons, access to individual sensors is generally not possible; rather, data wrappers are restricted to formal query binding patterns using a fixed set of filter parameters. This limitation entails a specific kind of cluster sampling. Additionally, very often the feasible number of access operations on a portal at a time is bounded for various reasons such as (i) the amount of requests assigned might cause an overload of the portal server system and/or may lead to service denial, (ii) the amount of data to harvest might turn out too time consuming, or (iii) the data carry little information suggesting a reduction of observations, or data are not available at all.

Naturally, sensor populations are stratified by the portals monitored. Now, seeking to strike a balance between good population coverage through iterated harvesting and a parsimonious use of portal access resources, an additional stratification of sensor populations based on price change dynamics [9] is introduced. To this end, an *active harvest heuristic* estimates the expected change rate, or latency, of a product expressed in terms of regular harvest intervals from both the frequency of observed price changes and the (relative) magnitude of change. Roughly, letting $p'_i = (\Delta_t)^{-1} \left| \frac{p_i - p_{i-1}}{p_{i-1}} \right|$, logistically transformed *relative* price changes

$$\varphi(p'_t) = \frac{\exp^{\beta(p'_t - \gamma)}}{1 + \exp^{\beta(p'_t - \gamma)}} \quad (1)$$

are averaged for estimating the expected latency

$$\tilde{\lambda} = \left[n \left[\sum_t \varphi(p'_t) w(\Delta_t) \right]^{-1} \right] \quad (2)$$

using the time lapse Δ_t between successive price observations p_{t-1}, p_t of the hitherto observed price series $p_0, p_1, p_2, \dots, p_t, \dots, p_n$ of this sensor as a weighting factor (with weighting function w monotonically decreasing in its argument). Then, $\tilde{\lambda}$ is converted to the active harvest weight

$$w_A = \Psi(1 - \Psi)^{\tilde{\lambda}} \quad (3)$$

of a sensor, based on some initially set inclusion probability $0 < \Psi \leq 1$. Accordingly, sensors exhibiting more frequent non-negligible price changes receive a higher active harvest weight (\propto probability) for inclusion in the upcoming harvest schedule. Actually, this active weight, adapting to the observation history of a sensor, is combined with a further *delay* weight (compensating sensors overdue for observation because of either recently failed access trials or not having randomly selected them in recent harvest cycles). Active and delay weights are recomputed (updated) every time a harvest iteration has taken place and combined to the *harvest weight* of a sensor.

The right-hand part of Figure 1 exhibits the heuristically estimated harvest times for the simulated price series by overlaying asterisks to the respective step functions; the heuristic starts working after 40 “observed” harvest cycles used for parameter estimation.

3.2 Dynamic Population Segmentation

Based on the current harvest weights—but regardless of the sensor assignments to portals—a sensor population is segmented into a (pre-defined) number of harvest strata, or “strips”, pooling sensors of (adaptively estimated) similar weight, interpreting the stratum centre (e.g., median weight) as sampling rate for randomly selecting sensors of the stratum as observation candidates of the upcoming harvest schedule.

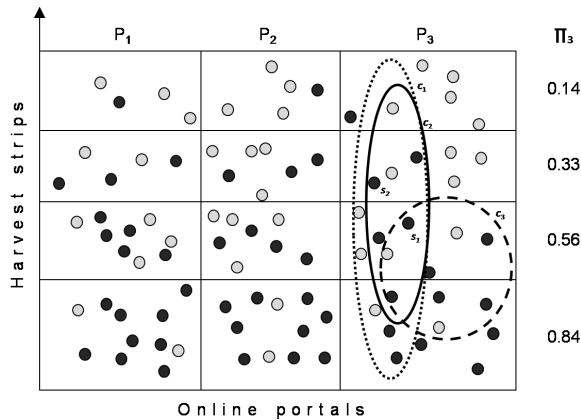


Figure 2: Sensor stratification over several online portals

Obviously, the effective size of the schedule depends on the respective current stratum sizes of the sensor population, the stratum samples are drawn from. Jointly with the implicit portal stratification, this weight-based segmentation induces a two-way stratification of a sensor population as sketched in Figure 2 (simplified to 3 portals and 4 harvest strips only), with randomly selected sensors in the preliminary pre-selection schedule marked black.

Most of the time, this pre-selection schedule is practically infeasible for actually harvesting online data since portal wrappers can access *certain* query binding patterns only each of which represents, in general, whole sensor classes. By logical necessity, sensor classes are always embedded in portal segments of a sensor population, but usually cut across harvest strips, as indicated to the right of Figure 2. Worse still, a sensor may entertain multiple class memberships (e.g., because of fuzzy assignments), classes may occur nested (e.g., any 4* hotel is also a 3* hotel by definition), and sensor classes often can be collapsed reasonably into larger sensor classes permitting less complex wrapper queries. Accordingly, the sampled sensors of the pre-selection schedule D have to be mapped into a suitable set of “wrappable” sensor classes to (i) the best degree possible such that (ii) all imposed portal access constraints are met.

4 Harvest Balancing

In what follows, let $s(c)$ denote the function returning the set of sensors in sensor class c and let J_m denote any subset of sensor classes available for building a covering set for D within portal m . Furthermore, assume that $J = \bigcup_m J_m$ still allows the identification of original elements in the J_m sets. Now, assume some real-valued finite cost bound $\chi_m > 0$ for each of the q portals relevant for a given sensor population instance, and let $y_m(c) \geq 0$, $1 \leq m \leq q$, denote the real-valued functions calculating (estimated) access costs of actually scanning the sensor class c on portal m .

4.1 Feasible Harvest Schedules

Using the terminology introduced and writing $|s|$ for the set cardinality of set s , the optimal feasible harvest schedule can be determined as solution of a “0-1 knapsack

problem" [14] as follows:

$$\text{find } \arg \max_J \left| \left(\bigcup_{c \in J} s(c) \right) \cap D \right| = \arg \max_J \sum_m \left| \left(\bigcup_{c \in J_m} s(c) \right) \cap D \right| \quad (4)$$

$$\text{subject to } \sum_{c \in J_m} y_m(c) \leq \chi_m, \text{ for } 1 \leq m \leq q. \quad (5)$$

Clearly, the identified solution set J^* of sensor classes may not provide a unique harvest schedule, particularly if $D \subseteq \left(\bigcup_{c \in J} s(c) \right)$.

4.2 Harvest Schedule Tuning

Having obtained a feasible harvest schedule of optimal D -coverage may still leave unsatisfied a couple of additional criteria. Arguably, it is advisable to compose the coverage of D of mutually *disjoint* sensor classes, even though overlaps do not conflict with resource constraints expressed as cost bounds of portals—clearly, the effort of successive processing of harvested data increases proportionally with the size of the data sets generated, regardless of their actual redundancy. Moreover, it is desirable to avoid querying the same offers several times in a single harvest cycle, as this could increase the threat of being recognized as an unsolicited source of Web server load which might cause access denial as a worst case. A further factor entering the objective function could represent a cumulated *valence* based on a non-negative function of the harvest weights of sensors comprised in sensor classes (such as an average, median, or mode value) emphasizing the inclusion of sensor classes contributing a larger share of sensors to sensor population strata with higher sampling rates.

Apparently, in view of the notorious complexity of knapsack problems, solutions exploiting suitably tailored meta-heuristics might work quite efficiently as a replacement of the standard dynamic programming approach.

5 Summary

Transferring market research into online contexts is quite challenging, in particular if one looks for a fairly automatic, application-independent methodology of online market monitoring. This paper has focused on the specific aspect of adaptively generating feasible randomized observation schemes—termed “harvest schedules”—aiming to allocate the data collection effort towards market segments exhibiting higher volatility (in terms of changes in offers as well as, in particular, prices of offers) as compared to apparently more stable market segments. The devised approach towards online data harvesting capitalizes on Web mining methods suggested in the literature, applied to the problem of tracking changes of Web sites or in documents accessible online. In doing so, the specific access conditions of online portals are taken into account such as (i) the re-identification of previously registered online offers (using record linkage techniques), (ii) the update of offer population registries, (iii) computing and maintaining adaptive weights associated with each offer tracked, reflecting the probability of re-harvesting an observation unit in the upcoming harvest schedule, and (iv) the derivation of harvest schedules matched to the constraints of online data access based on combinatorial optimization using adaptively segmented target (sensor) populations. Since routine statistical sampling methodologies do not apply straightforwardly, task-specific heuristics capturing the dynamics of market monitoring have been developed using the domain of (e-)tourism as a prototypical test bed of SEMAMO. The system operates in a cyclic mode, repeating over and over again the job sequence of data harvesting, data cleansing and rectification, weights updating, and adaptive harvest

schedule preparation based on the updated data gathered accumulatively in a historic database.

To date, in the SEMAMO project the processing framework, its main functional and storage components, and the mechanics of the harvest cycle have been developed and implemented prototypically. Parallel to system development, real online data in the domain of tourism are extracted to empirically evaluate the conceived heuristics governing adaptive data harvesting. Currently the proposed method of adaptive harvest scheduling is benchmarked against (i) non-adaptive data extraction schemes, and (ii) adaptive harvesting using a non-stratified random selection of observations, respectively.

References

- [1] C.-C. Wen and C.-C. Wen. The Use of Modern Online Techniques and Mechanisms in Market Research. *Proceedings of the CIMCA and IAWTIC*. IEEE Computer Society, page 93, Los Alamitos, US, 2006.
- [2] N. Walchhofer, K. A. Froeschl, B. Dippelreiter, M. Poettler and H. Werthner. SEMAMO - An Approach to Semantic Market Monitoring. *Journal of Information Technology and Tourism*, 2009, forthcoming. Already available at <http://www.ec3networks.at/semamo>
- [3] H. Werthner and S. Klein. Information Technology and Tourism - A Challenging Relationship. Springer Verlag, Wien, New York, 1999.
- [4] R. B. Doorenbos, O. Etzioni and D. S. Weld. A Scalable Comparison-Shopping Agent for the World-Wide Web. In *Proceedings of the First International Conference on Autonomous Agents*, pages 39–48, 1997.
- [5] G. Wiederhold. Mediators in the Architecture of Future Information Systems. *Computer*, 25:3:38–49, 1992.
- [6] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha and A. Crespo. Extracting Semistructured Information from the Web. *ACM Proceedings of the Workshop on Management of Semistructured Data*, pages 18–25, Tucson, US, 1997.
- [7] R. Baumgartner, O. Frölich, G. Gottlob, P. Harz, M. Herzog and P. Lehmann. Web Data Extraction for Business Intelligence: the Lixto Approach. In *Proceedings of BTW*, pages 48–65, 2005.
- [8] D. E. Appelt. Introduction to Information Extraction. *AI Communications*, 12:3:161–172, 1999.
- [9] J. Cho and H. Garcia-Molina. Estimating Frequency of Change. *ACM Transactions on Internet Technology*, 3:3:256–290, 2003.
- [10] C. Grimes and D. Ford. Estimation of Web Page Change Rates. *JSM'08: Proceedings of the international Joint Statistical Meeting*. Denver, US, 2008.
- [11] N. Matloff. Estimation of Internet File-access/Modification Rates from Indirect Data. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 15:3:233–253, 2005.
- [12] P. S. Levy and S. Lemenshow. Sampling of Populations: Methods and Applications. *John Wiley*, New York, US, 1999.
- [13] R. Baumgartner, O. Frölich and G. Gottlob. The Lixto Systems Applications in Business Intelligence and Semantic Web. *ESWC*, pages 16–26, 2007.
- [14] H. Kellerer, H. U. Pferschy and D. Pisinger. Knapsack Problems. *Springer Verlag*, 2005.