Semantic Online Tourism Market Monitoring

Norbert Walchhofer[a], Milan Hronský[a],
Michael Poettler[b], Robert Baumgartner,[c]
and Karl A. Froeschl[d]

[a]EC3 – Electronic Commerce Competence Centre, Austria
{norbert.walchhofer,milan.hronsky}@ec3.at

[b]Institute of Software Technology and Interactive Systems, EC Group
VUT Vienna Technical University, Austria
michael.poettler@ec.tuwien.ac.at

[c]Lixto Software GmbH, Austria
robert.baumgartner@lixto.com

[d]Dept. of Scientific Computing
University of Vienna, Austria
karl.anton.froeschl@univie.ac.at

Abstract

SEMAMO (SEmantic MArket MOnitoring) is a research project seeking to make use of the increasingly growing information available at Web-based sales and marketing channels for continuous market research. Assuming that online channels indeed mirror salient market developments faithfully, SEMAMO implements a nearly fully automatic adaptive data capture and analysis process delivering customer-defined market reports on demand. The paper describes the SEMAMO prototype implementation and exemplifies the functionality and utility of the approach in the domain of e-tourism, with a focus on the type of reports and visualisations the software, albeit not completely finished yet, can already deliver based on real-world data. Additionally, the role of formal domain description in SEMAMO is emphasized.

Keywords: online market intelligence, domain model, e-tourism, data aggregation

1 Introduction

In the Internet-based economy, traditional market research (Marder, 1997), including methods of market segmentation and price discrimination, no longer work the way they used to. In particular, the transparency of e-markets and the speed of market movements call for an increasingly comprehensive and efficient monitoring of markets and competitor behaviour (e.g., cf. Doorenbos et al., 1997). More specifically, in a tremendously competitive environment such as online tourism, the *continuous* observation of market behaviour is vital for every market participant:

offers vary more dynamically, prices are set more frequently, and are quite often changed even on the spot. As a consequence, e-markets are much more volatile and sensitive due to the relative simplicity to put such changes in place. Accordingly, e-businesses need a concise, recent and comprehensive picture of online markets to guide their product marketing strategy, particularly including pricing decisions.

Advanced information technologies – and particularly *semantic* technologies (Sheth & Ramakrishnan, 2003) – provide a means to intensify and accelerate market observation by reducing the cost of information procurement, thereby expanding the scope of competitive decision-making. Addressing this issue, the SEMAMO (SEmantic MArket MOnitoring) project, presented in this paper, explores an approach to continuous online market monitoring which extends current business intelligence methodologies and solutions by linking established statistical methods of market research to the mechanised collection of online market data. The SEMAMO system gathers business information available at Web-based sales and marketing channels, using semi-automated analysis processes driven by an (interchangeable) semantic model of the application domain.

Assuming that the Web indeed truly maps market developments in terms of product descriptions, distribution, promotions, and price developments in a timely and sufficiently accurate manner, SEMAMO implements a flexible modular framework of online market intelligence. The following sections, in turn, describe the main architecture components of SEMAMO (Sect. 2), highlight some of the market reports derivable form gathered online data (Sect. 3), and provide methodological details on the role of semantic technologies in domain modelling (Sect. 4). A brief summary of the SEMAMO project concludes the paper.

2 SEMAMO Architecture

SEMAMO monitors target markets, represented through a set of – pre-selected – Web portals, over time to detect changes in features declared relevant in a defined business context. To this end, a directed data flow from portals to customer-specified business reports is periodically cycled through, as shown in **Fig. 1**: The active data harvesting stage uses *Web wrappers* (Baumgartner et al., 2005) to collect data from portals – the SEMAMO sampling methods decide the optimal amount of data from the Web that is needed for a representative market picture without trying to be too exhaustive. Gathered data is *rectified* and prepared for subsequent processing and aggregation in a data transformation stage, and forwarded to the stage of statistical *analysis*, reporting and visualisation.

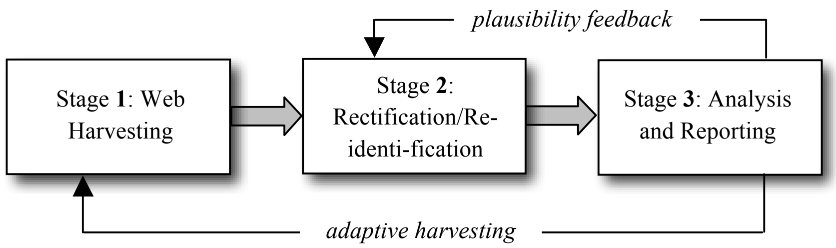

Fig. 1. Main SEMAMO processing schema

While this cardinal data flow is modelled in generic terms and thus basically application-independent, the semantics of a particular market monitoring application is captured in a separate *semantic domain model* component (cf. Sect. 4) also controlling the respective data processing flow. The transformed (validated and cleansed) data is accumulated in a permanent internal repository delivering aggregate information to both, adaptive process control and the preparation of markets reports. The data repository distinguishes between (i) *offer* data, recording time-dependent market features (in particular, offer price tags), and (ii) *registry* data, storing pivotal economic entities such as products (e.g., package tours, hotel rooms) and sellers/distributors (e.g., tour operators, hotels). Both types of data holdings are continuously updated based on cyclically harvested Web portal data, involving entity recognition methods for the accurate (though probabilistic) re-identification of entities already registered.

SEMAMO conceptualises monitored markets in terms of individual products, observed over time. **Fig. 2** illustrates the generic internal representation of (online) market structures, using an excerpt of an e-tourism application. While products are monitored, market aggregators or distributors actually offer them through one or more online portals. Accordingly, any product may in fact appear on several portals offered by the same or different sellers or intermediaries. This gives rise to a three-tier representation, consisting of a set of SAMPLER populations gathering products by type, a set of SWITCH populations gathering the product distributors or aggregators (intermediaries) as well as an artificial set of SENSOR populations. The latter is linking individual members of matched SWITCH and SAMPLER populations with one portal at a time, respectively, to represent a unique product offer named, for the sake of genericity, a *sensor*. Actually, each sensor inherits all defined and recorded attributes of the portal, switch and sampler instances it is a composition of.

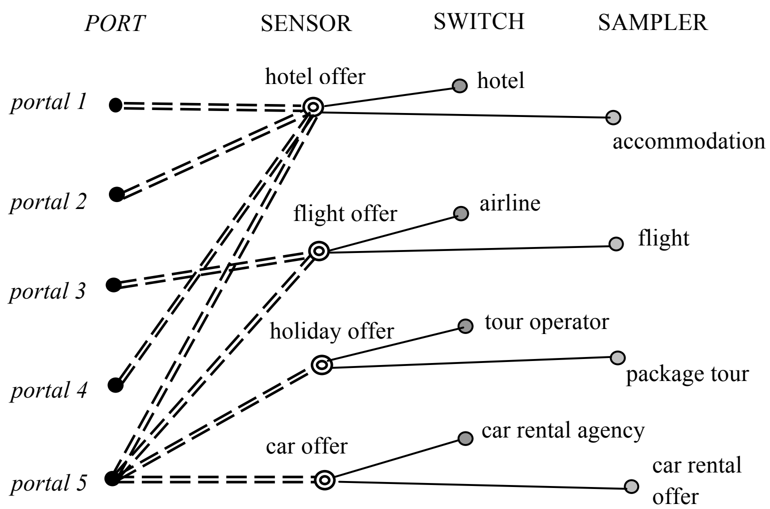

Fig. 2. SEMAMO market structure representation

SEMAMO operates cyclically; in each harvest cycle (e.g., every 6 hours), a subset of registered sensors is accessed online. Every time a sensor is accessed, or harvested, an observation value is generated (and stored) which is assumed to represent a valid price tag of the offer observed. Thus, repeated observation of a sensor generates a portal-dependent price histories for the offer it tracks.

In the online context, quite specific observation conditions prevail. Most critically, wrapping from Web portal induces server traffic which must not exceed certain levels, in order to prevent access denials to the wrapper demons. In general, sensors cannot be accessed individually but rather as whole classes of sensors as expressible by the query logic of the respective portals. Moreover, accessing a portal may not always produce the sensor data wished for. Finally, offer prices exhibit a more or less dynamic behaviour. SEMAMO seeks to account for these inherent peculiarities by an *adaptive* data harvesting strategy (cf. **Fig. 1**): based on previous evidence of harvest performance and, more importantly, the observed price dynamics of sensors, harvest samples are prepared allotting larger shares of the sensor population to market segments exhibiting more dynamic behaviour as opposed to less volatile market segments. On the individual sensor level, a sensor observing price changes more often or of larger magnitudes, is sampled more frequently compared to other sensors, and vice versa. Thus, while restricting the overall effort of online data procurement, market coverage and reporting timeliness can be improved substantially without compromising the statistical accuracy of analyses and forecasts.

The offer data repository of SEMAMO accumulates individual sensor price histories for statistical aggregation and analysis. Customer requests specifying certain market reports make use of this data holding by extracting relevant price histories. While possibly being processed on an ad hoc basis, it is assumed that, in general, monitoring requests entail a rather continuous observation of certain market segments, or features, to be reported on periodically. Reports may be composed either in a pull or push mode; pull requests specify a desired reporting periodicity while push requests generate reports triggered by predefined market events (e.g., price changes above a certain threshold). A set of typical reports that can be derived from SEMAMO offer data through pre-configured requests by choosing interactively from several statistical analysis, break-down, and visualisations options, is presented below.

3 Sample Market Reports: The Tourism Domain

As a trial application, SEMAMO uses the domain of hotel room prices, with a geographic focus on offers in Germany. Accordingly, the market structure reflects the top row of **Fig. 2**, with hotels in German cities and regions as instances of a SWITCH population "hotels" and hotel rooms to book as instances of the SAMPLER population "hotel accommodation". The Web portals used for data harvesting are concealed for presentation purposes, but the data shown are real, and collected during 2009. Obviously, the restriction to German-based hotel accommodations provides a fairly limited view of (global) tourism markets; yet, the example is considered comprehensive enough to highlight the virtues of both, adaptive harvesting and timely online reporting/visualisation of monitoring results.

In illustrating (a small selection of) the reporting capabilities of SEMAMO in terms of three representative, yet contrasting use cases, attention is restricted deliberately to the market reporting dimensions of price *levels*, price *variation*, and price *distribution*, broken down with respect to attributes such as

- time, also incorporating seasonality, seasons, workdays, or weekdays;
- geography, with its hierarchical political decomposition (such as country, state, and city) as well as subdivisions into tourism destinations, including the relationship between both political and tourism regions;
- other attributes defining an offer, like hotel ratings, number of beds in the hotel room/room type, and so on.

Clearly, it is also feasible to draw comparisons, with respect to any of dimensions mentioned, between portals as well as between distributors (intermediaries), in order to explore structural market properties and their dynamics, respectively. Actually, because of the built-in flexibility of the price histories selection functionality of SEMAMO, customers are fairly free to define "their" market segments of interest, provided these can be expressed in terms of available and recorded offer (that is, sensor) attributes. The data harvesting strategies can also be modified over time, based on the primary interests of customers, such as sampling interesting market segments more frequently. Data selections, then, may be combined flexibly with a wide range of statistical analysis methods, from simple data description through model-based price estimation (such as sector price indexes) to market segmentation and forecasting.

3.1 Use Case: Market Overview

Tourism stakeholders – e.g., hotel chains, tourism boards, or tourism institutes – have a natural interest in overall price levels in the hotel room market, its structural composition and development. For the sake of a specific illustration, assume a Bavarian tourism institute representing several hundred of members throughout the state. One of the institute's main tasks consists in providing a centralised marketing hub. Now, in order to make an informed pricing decision, all the recorded room bookings with the prices actually paid are demanded. Most hotels participate in international reservation systems displaying vacant rooms with their price tags, hence there are several (online) channels to book a room, very often at varying rates for each of the channels.

At this point, SEMAMO provides the required data by monitoring the (most) relevant Web portals the pertinent Bavarian hotels use as their preferred sales channels. **Fig. 3**, left-hand side, exhibits a chart of price levels of double rooms for all German states and nearly 200 of the largest cities, including state capitals (marked by squares). The actual price levels can be gleaned from the grey-shaded scale next to the price map, with the scale to the left referring to states, and the scale to the right to cities. The right-hand side of **Fig. 3** presents the price development of double room offers in Bavaria during summer of 2009, with the abscissa denoting the week of advance booking one month ahead of time. Besides the overall and Bavarian average price levels as juxtaposed to those of four major Bavarian cities, the dotted lines indicate the standard deviation of prices. By the way, in week 38 (mid-September), the city of Munich displays a markedly deviant price behaviour (at least for double rooms), signifying a major event of tourism relevance taking place.

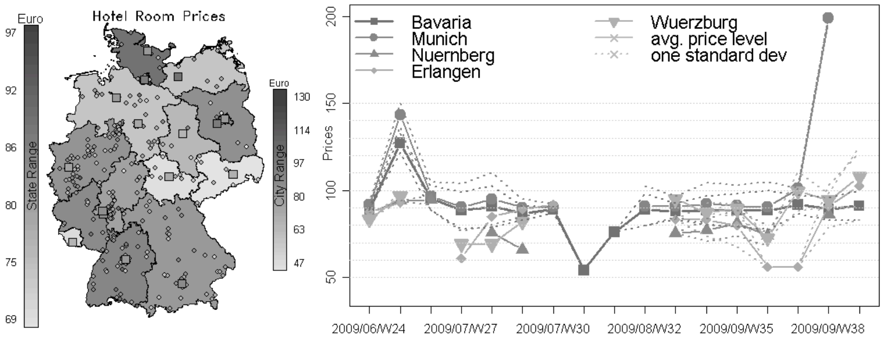

Fig. 3. Price map of German hotel room prices and their temporal development for Bavaria and four selected major Bavarian cities

Having at its disposal such kind of market information (which, by the way, can still be drilled down considerably), the considered tourism organisation is in a good position to adjust marketing decisions to both market state and dynamics. Furthermore, these reports can be carefully aligned to local specialties or events (such as folklore events, art festivals, etc.), or used to watch the performance of one's own, or the competitors', marketing campaigns. Apparently, by using such a market monitor, a tourism stakeholder cannot only observe its home market but has, in principle, equally easy access to the market data of other tourism destinations, the region is competing with.

3.2 Use Case: Peer Group Analysis

Another very interesting use case is about observing *micro markets* with only a few players acting in a tight regional competition. Often, peer group comparison is interesting to business entities to watch their strongest rivals.

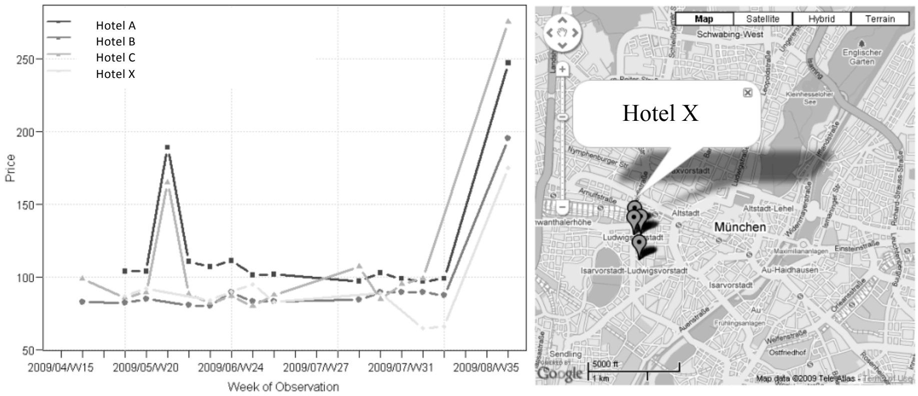

Fig. 4. Peer group analysis of four hotels located near the main station of Munich with a Google map view attached (prices one month before departure)

Hence, if a particular hotel located, say, in the centre of Munich, as illustrated through the Google Maps mashup in **Fig. 4**, is about to set a price of double rooms, there are several factors to account for, such as current room capacity, season, economic conditions, and price arrangements with cooperating business partners. Many of these can be dealt with by fairly standard internal business intelligence tools; however, the additional usage of competitive information on relevant peers as offered by SEMAMO contributes crucially to making well-informed business decisions.

In order to highlight the flexibility offered by SEMAMO, in **Fig. 4** an exemplary *peer group analysis* is shown for some 3* 'Hotel X' and three of its nearest neighbours. On the left-hand side of the exhibit, the price development for double rooms of these 3* hotels is depicted from June to October. As can easily be seen from the chart, 'Hotel X' is acting in the lower price segment, which makes it even more essential to know its competitors' prices. Assuming that the three "peer" hotels depicted are X's only local competitors, it becomes apparent that X might easily raise the price to about € 20 higher, from end of June to the beginning of August, without exceeding competitor offers. The steep rise in prices towards September (bookings for October), by the way, is reflecting the collective annual price adjustment anticipating the popular 'Oktoberfest' in Munich, as already mentioned in Sect.3.1. Furthermore, the chart illustrated in **Fig. 4**, if combined with geographical data, can easily boost a hotel's pricing decisions without doing any further market research on its own.

3.3 Use Case: Market Structure Analysis

The adaptive data harvesting process decides which offer (i.e., sensor) classes to observe how frequently. For harvesting, selected sensors are mapped to values of a deep web search on attached Web portals. As a result, a number of harvest records are returned for each such query. For the adaptive data harvesting procedure to operate properly, there are three phases to pass through: market exploration and calibration of offer behaviour as initial phases, followed by the actual continuous monitoring operation

Market Exploration Phase
As a first step, the target market, or market segments, have to be defined. In the tourism domain, the market characterisation includes (i) the geographical focus and (ii) several offer selection criteria, together defining the queries (i.e., search and filter parameters) and their instantiations submitted to the attached Web portal wrappers. Accumulated query responses – some of which may remain empty, even after several re-trials, and entail a cancellation of involved sensor classes – provide the *effective* coverage of the target market, with its accuracy and structural distribution depending on the overall market variability. Upon convergence of iterative market exploration, in this phase all accessible domain entities (such as hotels, hotel rooms on offer, etc.) have been extracted and registered.

Parameter Calibration Phase
After the exploration phase has generated an initial market coverage, sensor price histories are gathered. Since, in the beginning, neither the frequency nor the magnitude of price changes are known, another couple of harvest iterations is required to stabilise estimates. To this end, a *market census* including virtually all sensors collected in the market exploration phase, is taken repeatedly. Apparently, market census processing may encounter further domain entities not yet registered, e.g., because of constraints on Web server access quotas, implying a continuous update of the registry. Provided that estimates of price dynamics (price change rates) stabilise eventually, the initial parameter values necessary to run the routine adaptive data harvesting are available.

Adaptive Monitoring Phase
After calibration, the monitor enters the regular sequence of harvest cycles. In each cycle, a random set, or *harvest sample*, of sensor classes is chosen for sampling based on previous evidence adaptively governing the heuristic selection of sensors depending on the sensors' historic (recorded) price histories or, rather, the parameters estimated thereof in the calibration phase (cf. Sect. 4.2, below). During the adaptive monitoring phase, still new domain entities will be encountered, while others perhaps may vanish, entailing a continued update and maintenance of registry data.

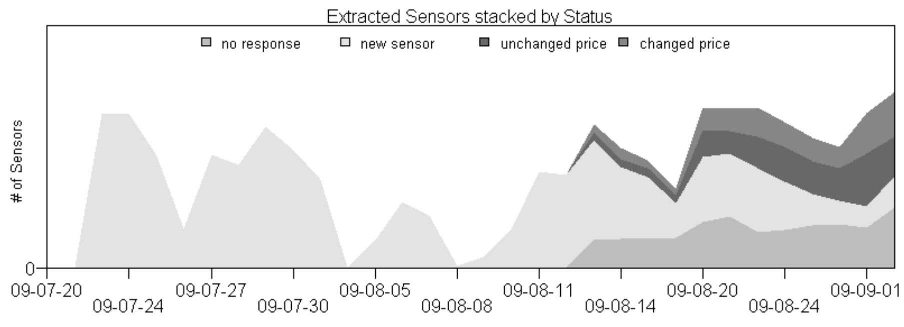

Fig. 5. Market exploration and beginning of the calibration phase of SEMAMO

The reports generated during these successive phases are, on the one hand, interesting for the SEMAMO operator to better understand the progress in each phase; on the other hand, they also shed light on general market dynamics, e.g., the frequency of price changes and of new offers occurring. **Fig. 5** illustrates the market exploration and calibration phases, and depicts the number of new hotel offers in each cycle as well as sensor classes that did not return any data. Moreover, on entering the calibration phase, for already registered sensors, both changed and unchanged prices are visualised. Based on this information, it becomes pretty obvious that the hotel market is highly dynamic, as about one half of the prices actually change between cycles.

4 Configuration by Semantics, and Adaptive Processing

In SEMAMO, the *automation* of data collection, statistical analysis, and market report generation processes is emphasized while a maximum of flexibility and general applicability of the core system should still be preserved. Thus, by logical necessity, the core system abstracts the contingencies of specific market monitoring applications, apportioning the specification of individual application properties to a dedicated configuration layer, called the *semantic domain model* (SDM, for short) of a monitoring application. The SDM specifications comprise, among configurations of minor relevance:

- the wrappers tapping Web portals of interest;
- SWITCH, SAMPLER, and SENSOR populations characterising the application domain (cf. **Fig.2**), including the variables to observe with value domains attached, etc.;
- the coupling of SENSOR populations to portal wrappers through virtual data *sources* abstracting the particular query logic of the wrapped portals;
- analytical markets, representing excerpts of the offer data repository used for statistical analyses and report generation, alongside with their dimensional structure for roll-up and grouping (that is, data warehousing; cf. Kimball & Ross, 2002) operations;
- report and order structures governing content and periodicity, or triggering conditions, of reports to generate;
- mapping functions converting (i) raw harvest data into internal, regularised offer data structures, (ii) the latter into analytical market data structures, and (iii) sensor classes to queries on Web portals.

Particularly data rectification processes resist simple standardisation efforts, entailing almost inevitably an individual configuration of portal wrappers, data cleansing routines, and entity recognition (as part of record linkage) methods; in this regard, little more than fairly comprehensive function libraries can be provided to support the composition of processing sequences.

4.1 Ontology-based Application Modelling

In favour of a modular set-up, the SDM of SEMAMO splits into two representation layers, viz. a (single and unique) *system* ontology, and a *domain* ontology depending on the respective SEMAMO application. In terms of a *domain structure description*, the elements of the domain ontology comprise the specific data models used by SEMAMO sources (i.e., portals) and registries, both of which linked to the structures predefined in the system ontology in order to make the general data harvesting and order processing logic applicable to the the respective application domain instances. To this end, both ontology layers are represented in OWL-DL (W3C, 2004) using SPARQL (W3C, 2008) as query language.

For example, in the hotel price monitoring case, referring to **Fig. 2**, a hotel data schema comprising attributes relevant for describing individual hotels becomes linked to the 'Switch' class of the system ontology; likewise, an accommodation data schema could be chosen for assignment to its 'Sampler' class. Ideally, structurally compliant domain ontologies – such as the *ebSemantics* accommodation ontology (http:// www.ebsemantics.net/doc/acco.owl [Sept. 4, 2009]) – can be utilised to simply "plug-in" established models into the SEMAMO system ontology. However, it must be remarked that standard domain models (such as the ones of the Open Travel Alliance; http://xml.coverpages.org/openTravel.html [Sept. 7, 2009]) generally do not specifically fit analytical purposes such as market research. For example, price quotes for the very same hotel room may (and will) differ if booked the day before, the month before, or the season before – but, in general, in the standards there will not be such a room attribute accounting for different booking leads.

Actually, in linking a variety of online data sources to a single central data model, a schema integration task (Ziegler & Dittrich, 2004) has to be solved effectively. As a consequence, in addition to the central domain ontology of a market monitoring application, providing a hub for the harmonisation of source data, the partial source data models linked to individual Web portals need to be mapped to the respective domain ontology. Since Web portals rarely publish their internal data models, even if these comply to one of the established model or data exchange standards, both the domain ontology and the associated source mappings of an application have to be handcrafted most of the time before they can be inserted into the SDM and connected to the respective system ontology elements. Once arranged, the mappings feed harvest data from accessed portals automatically to the successive SEMAMO processing stages.

A second area of the SDM concerns *domain processing data*, i.e. domain-dependent information controlling the processing of harvest data. In brief, stage 2 of the SEMAMO processing schema (cf. **Fig. 1**) can be conceived as a compound function

$$reid_D \; \bigcup_p reg_p \;\; ret_p \;\; Q_p$$

where

- p denotes a portal attached to SEMAMO;
- Q_p denotes a set of queries (i.e. instantiated binding patterns) submitted to portal p, i.e. $Q_p \quad dom \; A_1 \;\; \ldots \;\; dom \; A_k$, letting $dom \; A_j$ denote the value domain of the *j*-th attribute of the query space of portal p ;

- ret_p denotes the retrieval function returning a harvest sample, including offer prices, according to the source data model of portal p;
- reg_p denotes the function which regularises harvest samples towards the central sensor data model (applying schema transformations, data cleansing operations and value pre-processing possibly using additional lookup up data such as thesauri etc., or depend on distributional properties of entity populations, entailing yet another information link back from the statistical analysis stage of a SEMAMO instance; cf. **Fig. 1**), including the split of harvest records into sensor data and sensor price history data;
- the union of all regularised harvest data is pooled across all portals involved in a central transient *staging area*;
- D denotes a set of domain dependent discriminating rules used for sensor re-identification [➔ ev. In PMML repräsentiert?], derived from a domain-specific learning sample; and, finally,
- $reid_D$ denotes the function used for partitioning staging data into individual sensors matched against the sensor registry, and updating the query/sensor relation, respectively (cf. **Fig. 6**).

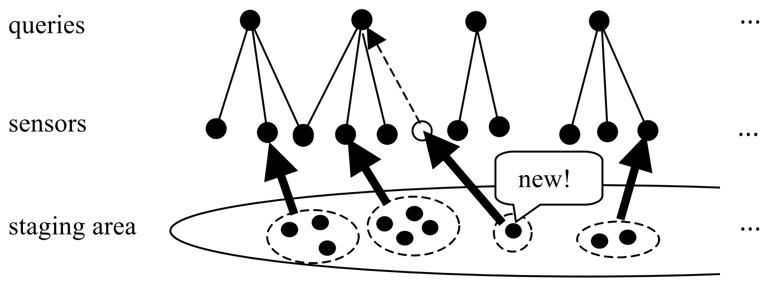

Fig. 6. SEMAMO query/sensor relation

4.2 Feedback Processes

By means of SDM configuration (and some additional manual tuning), a SEMAMO application instance, including all internal processes, is arranged. These processes are designed *adaptive* – cf. **Fig. 1** – in order to optimise resource utilisation and update the registry data repository (cf. Sect. 2).

As a pivotal design consideration of SEMAMO, data harvesting exploits previously gathered market information to adapt sampling schemes iteratively: price histories of individual sensors are converted to sampling weights by a specifically developed heuristic decision rule (Walchhofer et al., 2009), using the calibration phase (cf. Sect. 3.3) for parameter tuning. The derived sampling weights reflect the dynamics of different market segments and allocate sample shares accordingly.

During the market exploration phases, but occasionally also later on, the iterative data harvesting encounters domain entities – e.g., cities, or hotels, etc., in the running tourism example – not yet registered and, thus, left unassigned by the $reid_D$-function (cf. **Fig. 6**): every time a harvested entity cannot be re-identified by SEMAMO, a service process is triggered which updates the query/sensor relation (sometimes asking for the operator's assistance, though). This way, as a side effect, the SDM incrementally gathers also application-relevant knowledge such as, e.g., the link of cities to tourism destinations other than predefined political regions.

5 Summary

The SEMAMO project pursues research in online market intelligence by coupling methods of online data collection from Web-based sales channels with statistical tools of market research to generate timely and accurate aggregate information about online markets and their dynamics. Based on semantically rich models of monitored markets using dedicated representation standards and tools, a generic framework for data capture, storage, analysis, and reporting according to customer orders has been conceived and prototypically implemented. Moreover, the statistical functionality of SEMAMO has been designed adaptive in order to optimise the market monitoring effort without compromising market coverage, data quality, and accuracy. The SEMAMO solution is composed modularly, so that it is applicable quite easily to different domains by exchanging the formal representations of the respective market models.

As a test case, SEMAMO is applied to the domain of online hotel booking portals. Because of its rather complex product and service structures, the tourism domain provides an ideal candidate for proving the benefits of semantic technologies (Staab & Werthner, 2004). In this paper, use cases and sample reports drawn from this particular domain highlight (some of) the capabilities of the system to tourism marketers and decision makers whereas salient technical issues were indicated only in brief.

References

Baumgartner, R., Frölich, O., Gottlob, G., Harz, P., Herzog, M., & Lehmann, P. (2005) Web data extraction for business intelligence: the Lixto approach. In: Proc. BTW, 48-65.

Doorenbos, R.B., Etzioni, O., & Weld, D.S. (1997) A scalable comparison-shopping agent for the World-Wide-Web. In: Proc. AAMAS, 39-48.

Kimball, R. & Ross, M. (2002) The data warehouse toolkit: the complete guide to dimensional modeling. Wiley & Sons, 2nd ed.

Marder, E. (1997) The laws of choice – predicting customer behavior. The Free Press/Simon and Schuster.

Sheth, A.P., & Ramakrishnan, C. (2003) Semantic (web) technology in action: ontology driven information systems for search, integration and analysis. IEEE Data Engineering Bulletin, Vol. 26, No. 4.
Staab, S. & Werthner, H. (2002) Intelligent systems for tourism – Introduction. IEEE Intelligent Systems 17/6, 53-55.
W3C (2004) OWL Web ontology language use cases and requirements. W3C Recommendation. http://www.w3.org/TR/webont-req/ [Sept. 4, 2009]
W3C (2008) SPARQL query language for RDF. W3C Recommendation. http://www.w3.org/TR/rdf-sparql-query/ [Sept. 4, 2009]
Walchhofer, N., Froeschl, K.A., Hronský, M., & Hornik, K. (2009) Dynamic population segmentation in online market monitoring. Proc. IFCS 2009 (Dresden, Germany), revised paper submitted.
Ziegler, P. & Dittrich K.R. (2004) Three decades of data integration – all problems solved ? In: Proc. WCC 2004, 3-12.

Acknowledgements

The SEMAMO project is supported by grant Fit-IT "Semantic Systems and Services" No. 815.135 of the Austrian Research Promotion Agency (FFG).