SEMAMO in a Nutshell
A Brief “High-level” Description


Introduction

This treatise summarizes the “big picture” of the SEMAMO model, omitting any details of either concept or project. Instead, the purpose of this summary consists in establishing a converging view upon the SEMAMO project and its intended outcome amongst the participating researchers, partners, and stakeholders.

GENERAL CONSIDERATIONS AND REMARKS

First and foremost, SEMAMO is conceived as a tool for continuously scanning a certain part of reality, that is, SEMAMO gathers online data (using liXto wrapper technology) from relevant web sites in real-time, and memorizes regularized images of such scans for further processing.

As an empirical tool, SEMAMO seeks to exploit – presumed – regularities in the data collected. Hence, SEMAMO can be considered a statistical tool, yet a very specific one, as it is tuned to a specific kind of data generating processes entailing specific observation conditions. It holds as a fundamental assumption that any data can be gathered only once, as the observed data generating processes may never repeat their behaviour, that is, stationarity merely may be a testable hypothesis. As a consequence, observation data not captured for whatever reason at a time simply cannot be accessed or reproduced later on. Even if stationarity might be the true state of the nature, it normally takes time to get enough empirical evidence of this being so, entailing a series of successive observations. Therefore, SEMAMO relies on a memory of empirical states so observed.

SEMAMO is conceived as a tool implying quite a high degree of flexibility in fitting the system to a variety of application scenarios. To be economically reasonable, the eventual degree of flexibility needs to be decided, most probably based on experimental evidence gathered during the project phase. At any rate, flexibility presupposes a component structure of the system such that both the components themselves but also the alignment of components is open to configuration through various controls.

A salient precondition of the SEMAMO approach consists in the free accessibility of online data. Clearly, the envisaged service is feasible only so long as the commercial utilization of public data is lawful indeed. Resistance – to the point of access denial – from data publishers is to be expected, though, if SEMAMO operations interfere with the ordinary online business of web site owners, as this could (and most probably will) be deemed as a deliberate or grossly negligent activity affecting business adversely, or distorting competition, against which the offended parties may seek financial compensation or demand forbearance through legal enforcement.
RATIONALE

The overarching rationale of SEMAMO is of economic nature: given resources, as described below, should be allocated in a most profitable way. Resources consist in the effort of observation, undertaken for gaining data of interest such that statistical aggregates can be derived with demanded precision. Operationally, SEMAMO should support decision makers in striking an optimal balance in assigning available resources to observe different data sources in pursuit of maximizing (monetary) profits, given an opportunity to sell derived statistical aggregates (that is, market information). Economically, SEMAMO occupies the position of an intermediary adding value to existing market information by systematically collecting market information through continuous monitoring of online product/service offerings, and integrating these data according to defined (marketable) information requests. Thus, SEMAMO exploits the arbitrage from dealing wholesale in market monitoring, much in the way traditional market researchers do.

In order to economize its operation, SEMAMO needs to convert market opportunities – that is, the opportunities to sell market information to (business) customers – into a commensurable scale for the optimisation calculus. Presumably, as with any other commodity, there is some economic relationship between the price and the quality of the market information delivered, although the elasticity of this relationship may vary considerably from case to case. Furthermore, with respect to statistical information, it could be assumed that information quality amounts to precision as expressed in terms of variances, or standard errors of estimates derived from observed data. The benefit of such a view results from establishing a direct link between target quality and resource allocation: higher (statistical) precision correlates, ceteris paribus, with more effort devoted to observation, that is, larger sample sizes. In other words, the underlying economic calculus amounts to start from precision requirements (of delivered statistics), after weighing them according to their margin profits, to assign limited observation resources to accessible market data feeds.

Practically, this straight reasoning is compromised in various respects. For one thing, different marketable statistics may take recourse to overlapping source data, thus effectively reducing the observation effort by exploiting synergies. Conversely, however, gathering data may be impeded either by sparse populations (that is, too few observations with regard to the aspired precision levels are possible for substantive reasons), extremely high observation cost (e.g., because of risking to be detected in conducting extensive market monitoring resulting, in a worst case, in service denials, or losing access to a data source in its entirety), or technical restriction of data access limiting the observation quantity or frequency. Finally, data in themselves may possess an inherently poor quality because of, for instance, badly maintained web sites, or the like.

ARCHITECTURE

Acknowledging the principal optimising approach of SEMAMO (that is, maximizing economic benefit from statistics through assignment of online data sampling resources), and keeping in mind the desired tool flexibility of SEMAMO, a two-layered system architecture seems adequate:

- on an upper layer, the resulting system is defined in terms of a configuration specifying the structure of data collection and processing (out of a set of parameterized components strung together suitably);
• on a lower layer, the configured system operates in an adaptive way, driven by requests derived from marketed information products and adapting these dynamically to the ongoing data observation process.

The flexibility designed into SEMAMO thus separates a structural configuration, and an operational adaptation. While configuration necessarily has to be accomplished manually by specifying various system parameters such as domain data models, market definitions, product taxonomies, data cleansing routines, analysis procedures, reporting styles as well as a range of general process controls, adaptation amounts to a data- and demand-driven autonomous control of operation of the system once configured.

The configuration of a SEMAMO instance involves at least three dimensions of a configuration space captured in terms of a “semantic domain model”, viz.

• product configuration: this refers to the characteristics of commodities or services directly wrapped from the online data sources sampled, comprising both quantitative and qualitative description elements characterising the products monitored, including data elements required for establishing product identities (that is, for deciding, whether or not two observations actually refer to one and the same product);
• market configuration: contrary to product configuration, market configuration deals with what forms groups of products entertaining – from a consumer’s point of view, at least – some economic substitution relation – while some product features may not be relevant from a market perspective, the very same product could participate in different markets, though, entailing additional description elements of products distinguishing between markets;
• analysis configuration: this part of the SEMAMO configuration determines the range of analytical outputs (statistical aggregates) and, thus, data analyses, of the gathered observation data, and from which marketable SEMAMO services (or “reports”) are eventually generated on demand.

Despite the (proposed) suggestive naming of these configuration dimensions, it must be kept in mind that, in each case, both data models and procedures are in fact specified, as the definition of data models implies particular processing steps using or generating specific data structures. For example, configuring products entails the specification of wrapper components actually harvesting online data, whereas the choice of certain statistical analysis techniques determines the structure of aggregate data (statistical indicators) deposited in SEMAMO’s data warehouse. Likewise, market configuration delineates the structure of orthodata by linking canonical harvest data to analytical output requirements.

In a rough way, the overall data flow in SEMAMO is sketched in the chart below. To the left, the actual data sources (harvested portals) are indicated which are accessed through a tripartite structure composed of the tiers (top to bottom) of “ports”, “switches”, and “samplers.” The latter actually mark the logical products of which a variety of attributes is sampled. Switches simply go in between samplers (products) and ports (web portals) as, occasionally, products may be distributed to several portals through intermediaries (online merchants, typically). Depending on portal idiosyncrasies, the canonical data harvestable may take differing shapes; in part, this is mandated also in the more general case that a portal may serve several markets of different (formal) structure at the same time. Because of linking

---

1 To simplify reference, henceforth ‘product’ is the termed to denote either of commodity or service.
2 A typical example in this respect, hotel rooms – regardless of their inherent product features such as number of beds, services included, or hotel classification – take part in markets such as package tours, leisure vs. business travel, last minute booking, etc.
3 Many thanks to Norbert Walchhofer for producing the typeset version.
canonical data to portals individually, many harvest-specific details of canonical data reflect portal, or access, structures (notably, attributes, and terminologies used). Canonical data then is mapped to orthodata, reflecting the structure of target markets (in terms of external analyses). Orthodata and DWH structures are tightly aligned, of course, in that all data cube dimensions have to be supported by orthodata elements appropriately. Finally, reports are generated from DWH data. Technically, and for obvious reasons, the mapping of canonical data to orthodata must be reversible, because requested reports get re-translated into orthodata triggering further data harvest jobs upstream: after determining the portals relevant in processing an external analysis, canonical data (unless already available) must be procured such that the respective portion of orthodata is covered by canonical data to be harvested. Although not shown in the chart, canonical data bear a more detailed structure, distinguishing (at least) between stable product features and time dependent data elements (notably, product price).

Contrary to the static set-up of a SEMAMO configuration, the adaptation layer is responsible for the dynamics of data processing, given a defined configuration. Plainly, the configuration specifies a processing schedule with a few degrees of freedom subject to data- and demand-driven control, that is, self-adaptation. Analytically, the following distinction ensues:

- By design, the primary source of adaptation results from observed changes in product variability. Again, this can be broken down into changes in (i) product quality, or (ii) product price. Changes in product quality affect the decision about product identity, and thus needs a treatment different from changes in product price which is the observable monitored first and foremost, of course. Methodologically, prices are considered measurements taken from a (smooth) stochastic process while quality changes are “discontinuous” events, notwithstanding the fact that also prices are in fact set from time to time only, yet these steps still take place within a homogenous dimension. At any rate, changes in either quality of price or products have to be tracked, and the more
frequent so, the more such changes taking place are factually observed. May this type of adaptation be termed “harvest-level adaptation.”

- While keeping the SEMAMO configuration constant, different information requests may be served on demand. More precisely, a SEMAMO configuration will certainly reflect an expected, or experienced, demand mix on the longer term. Yet this does in no way rule out the possibility that the configuration permits a “margin business” by serving additional information requests within the boundaries of adaptation a particular configuration supports. This may happen to the degree, a SEMAMO configuration generalizes a set of anticipated requests, encompassing certain product markets “in their entirety.” Still, changing request patterns will, in general, affect the allocation of sampling resources due to the different precision requirements imposed as determined by the information quality (precision) sold to the requests’ customers. This type of adaptation may be termed “demand-level adaptation.”

Eventually, both harvest-level and demand-level adaptations merge in a dynamic harvest schedule, reflecting the currently optimal data observation scheme given the sensed data variability and request collection to be served.

Every now and then, the configuration of a SEMAMO instance has to be changed, raising the question of how to accommodate to such discontinuities. In the worst case conceivable, of course, the change is so profound that recorded and processed market data lose almost all of their value; then, a complete restart ensues necessarily. Otherwise, it might be reasonable to invest into internal schema transformations maintaining continuity in usage of memorized market data at least for some parts of the data stock.

**CORE PROCESS OPERATION**

The functional core process of a configured SEMAMO instance can be split into 5 processing streams running concurrently. Additionally, these streams could be partitioned into a repeated sequence of processing cycles, with the duration of which depending on the maximum cycle time\(^4\) of any of the five streams.

The five processing streams are as follows:

- **Job dispatch**: in this stream, both standing and additional requests are gathered, weighed according to economic benefit (that is, ranked according to the revenue/precision ratio), and inserted into a job schedule. Basically, the jobs scheduled define which data sources have to be sampled in which frequency unless the requested information products cannot be derived directly from the SEMAMO data warehouse or already available orthodata (that is, by reusing already harvested and processed data).

- **Dynamic harvest scheduling**: for the (current) harvest cycle, given a batch of jobs to be processed, the actual demand for data to be harvested can be determined (combining the requirements of all the jobs scheduled in the batch). Taking into account imposed limits on data access at wrapper level, the resulting harvest schedule determines the loads and accesses to the different data sources available.

- **Harvest**: executing the generated harvest schedule, this stream actually produces canonical SEMAMO data by wrapping the data sources (online servers) as indicated, taking account of deviations from the harvest schedule as good as possible; note that

---

\(^4\) Aka as „harvest cycle (time)”
this stream also keeps track of changes (through aligned service processes) in registries maintaining lists of servers and products.

- **Cycle (or, harvest) analysis**: reading in the outputs from the harvest process, this stream cleanses, rectifies, and thus generates SEMAMO orthodata ready for statistical aggregation, and deposits statistical results in the SEMAMO data warehouse (“storehouse”) as mandated by job requirements. As a by-product, both data rectification and data aggregation stages generate auxiliary control information to be fed back to the dynamic harvest scheduling stream (aka “internal analysis”), so that scheduling can properly take account of detected changes on the harvest level.

- **Delivery**: this stream, finally, converts accumulated (external) analysis results from the SEMAMO storehouse into (numerical) reports for finishing according to the customers’ orders.

Besides this core process, of course, various maintenance processes have to be run which, for the sake of brevity, are omitted from discussion for the time being.

The basic SEMAMO processing cycle can be split into 4 quarters distributed over the 5 processing streams as follows:

|                | I | II | III | IV | I | II | III | IV | ...
|----------------|---|----|-----|----|---|----|-----|----|------
| job management |   |    |     |    |   |    |     |    |      
| dynamic harvest scheduling |   |    |     |    |   |    |     |    |      
| harvesting     |   |    |     |    |   |    |     |    |      
| charge analysis|   |    |     |    |   |    |     |    |      
| delivery       |   |    |     |    |   |    |     |    |      

Thus, each job batch is executed in successive 5 quarters of a base cycle of the core process. Apparently, the ‘delivery’ stream, unless relying on the latest data harvested and analysed, may draw upon any harvest charges processed before; so, in case that a request does not entail the harvesting of “fresh data” (which, however, is only revealed through checking the corresponding job’s requirements against available storehouse contents, or orthodata, respectively), job processing may be finished already in as few as two successive quarter cycles.

From a resource management point view, it is of course advisable to achieve an approximately equal duration of execution time in each processing stream (by appropriate assigning of computing resources). Note that, obviously, the scheme of the core process presented here is by no means the sole one possible. Note also that by varying the harvest cycle time yet another general parameter is available for controlling the core process.

***