Stéphane Marchand-Maillet and Birgit Hofreiter

Big Data Management and Analysis for Business Informatics

A Survey

Modern communication networks have fueled the creation of massive volumes of data that may be valued as relevant information for business activities. In this paper, we review technologies for enabling and empowering business activities, leveraging the content of this big data. We distinguish between data- and user-related technologies, and study the parallel brought by the overlap of these categories. We show how the trend of Big Data is related to data security and user privacy. We then investigate automated ways of performing data analysis for Business Intelligence. We finally review how groups of users may be seen as a workforce in business through the notion of human computation or crowdsourcing, associated with the notions of trust and reputation. We conclude by discussing emerging trends in the domain.

1 Introduction

Progress in *Business Informatics* aim to develop business administration using computational and information technologies. As such, business informatics may use any method providing a technology useful for its end purpose.

Modern business activities essentially rely on an accurate management of knowledge (often referred to as *Business Intelligence*). The development of communication technologies and the wide-spread and ubiquity of communication networks have created an opportunity for gathering and analysing data in view of deriving useful knowledge. Hence, business informatics is primarily supported by data management and data analysis technologies. In addition, users and user groups remain at the center of any business. They may assist performing data analysis as much as benefiting from it.

In this paper, we review and analyse the main enabling technologies in business informatics. We explore as thoroughly as possible the information landscape in which business informatics operates, to understand the aspects and their characteristics, potential risks and benefits. User-generated data is considered a potentially rich source of information for business and user behaviors are modeled using this data. Users and data are therefore two inter-related main *actors* within this landscape that we explore via these two perspectives, as illustrated in Fig. 1.

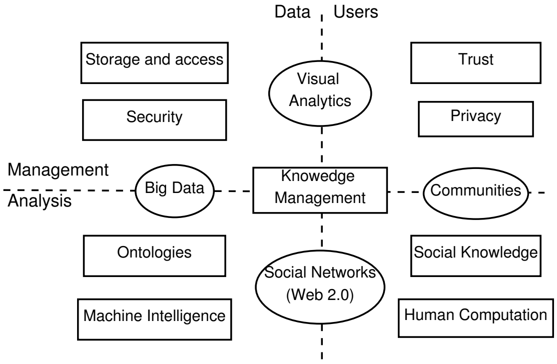

Figure 1: A classification of domains for enabling and empowering technologies in business informatics. Square boxes indicate technical domains, whereas circular items relate to multi-disciplinary binding fields of study

Investigating data-related aspects allows to understand the technical infrastructure that should be set up and sustained, from base data collection and housing to sophisticated data analysis. In parallel, studying user-related issues allows to model the user and his community, as originator of this data. We therefore distinguish between data- and user-related technologies, although the split is somewhat artificial since these technologies generally overlap. We further look into passive *management* technologies (whose main aim is not to create any knowledge) against active *analysis* technologies (that transform data into knowledge). The extend of our review is symbolised in Fig. 1. Section 2 reviews the housing and preservation of data, in relation to the current challenge of *Big Data.* In Sect. 3, we review automated technologies for data analysis. These are crucial for their aspect of scalability, since any necessity for user intervention would create prohibitive costs at large scale. As a mirror to the data-related sections, Sect. 4 reviews user-related strategies, from preserving user privacy to exploiting the force and intelligence of the crowd. We discuss the future potential of reviewed technologies and foreseeable extensions in Sect. 5.

2 Data management

While information and communication networks have triggered the creation of an overwhelming mass of data, they have also created opportunities to monitor and mine this data, thus augmenting drastically the volume of contextual information potentially available. Massive collections of digital documents are made available, either publicly on the Web or in private networks related to companies, workgroups or social associations. Data may be very diverse and arise from any form of information exchange. Examples of this include:

- Textual: Web pages (personal, professional, from individuals, groups or companies), emails, blogs posts, news feeds, exchanges over social networks;
- Multimedia: photos, videos, music, audiovisual blogs, ...;
- Process data: sensing data (GPS, weather, traffic monitoring, ...), computation (sociological, scientific, financial trends, ...);
- Logs: traces of human interaction with systems (information, e-commerce, entertainment, ...), logs of machine-machine communications (web services, distributed computing, ...).

Before considering data analysis, a choice is to be made on the form of data housing, if any. The emergent paradigm of big data (Sect. 2.1) is addressing some choices there. In turn, data preservation and access immediately open security issues, reviewed in Sect. 2.2.

2.1 Big data

Every study on the topic shows clearly that the volume of data created by individuals and companies is growing exponentially (see, e.g., Manyika et al. 2011). In parallel, analysts predict success to anyone who will exploit this data accurately, thus implicitly supporting the meaningfulness of this data. However, this data is everything unlike what companies are used to deal with. It is unstructured and redundant and potentially noisy or corrupted. Every piece of the data may be seen as noise that would pollute a database of clean and structured data. There is nevertheless a clear intuition that the mass compensates for the defects of the pieces. A global picture of the data should contain information that could be exploited to many ends. This is the challenge of the recent trend identified as *big data* (Big Data: Science in the Petabyte Era 2008; ERCIM News Special Theme: Big Data 2012).

Big data has initially been characterised by its "three Vs" (Fayyad 2012), mostly addressing its technical specificity:

- **V***olume*: In itself, the volume of this data is an issue. It surpasses many of the simple storage strategies classically used. At this scale and evolution rate, it is hardly possible to structure and clean the data, for both technical and cost reasons;

- **V***ariety*: In order to transform data into knowledge, its multiple facets should be taken into account (see Sect. 3). The data in question therefore encompasses a high diversity in its content, format, structure and interpretation. Again, this goes against most principles of classical data management and storage paradigms where the structure of the data should be understood and stable;
- **V***elocity*: One of the main characteristics of big data is the pace at which it is generated and at which it evolves and gets obsolete. In other words, this data inherently bears a strong temporal dimension. Usage logs, trends, news, are all content that have a strong interest in their immediate history and quickly decline into useless or even polluting data. However, this data may also have a behavioral interest at long-term on a more global temporal scale (e.g., Morrison et al. 2012).

These technical factors prevent the use of the typical database models (e.g., a relational structure made usable via SQL) and impose to move onto more flexible, agile and scalable structures (including the *NoSQL* trend[1] advocating for schema-free storage or the MapReduce model (Dean and Ghemawat 2004; Mohamed and Marchand-Maillet 2012)[2] to support indexing, e.g., via the agnostic name-value pair model). Decentralised storage and processing systems (a.k.a Peer-to-Peer systems or *the Cloud*) rely on structures having these characteristics to make the data safe, ubiquitous, and accessible ("Anything, Anywhere, Anytime").

In practice, the current appeal for big data has mapped into new functions coined as *data scientists*, i.e, data analysts able not only to perform technical operations on the data such as preparation, cleaning, compaction, *etc*, but also to contextualise the data and read it with all surrounding parameters (e.g., social context, as detailed in Sect. 4.3).

Many "Vs" have been added to big data (e.g., Viability, Veracity (Vossen 2013), Volatility, ...) but the main "V" business is concerned with is

- **V***alue*: The question is "how to make value out of this large, complex and unstable stream of data?"

There are many answers to that question, including:

- By better understanding actions and behavior of its customers, traced by the log of their actions, a company will be able to offer better and more relevant services;
- By better understanding the context within which it operates, characterised by the mining of environmental factors, a company will lower its risks.

The first common step is always to transform data into knowledge, partly thanks to the technologies reviewed in the next sections. One finds reports on success stories of big data analytics[3] relating how such insurance company could fine-tune its risk model using deep data analysis, including exploiting inferred customer social relationships (which in turn poses questions on privacy - see Sect. 4.1. Looking at business as a permanent complex constrained optimisation problem, where the right balance should be found ("price *vs* volume, cost of inventory *vs* the chance of a stock-outrisk"[3], *etc*), the success of big data is in providing insights into how to rationalise decisions on these tradeoffs. In that case, the *noise* in the data refers to any potential inconsistency in data patterns, unintentional (e.g., failures in process or communication), or intentional (e.g., spam (Mukherjee et al. 2012)).

It should be emphasised there that as much as there may be some benefit for a company to extract knowledge from the data, there is an inherent risk that this data is used by an adversarial party in many ways. These includes industrial

[1] Web-scale databases: http://nosql-database.org

[2] The Apache Hadoop library: http://hadoop.apache.org/

[3] e.g., http://www.forbes.com/sites/mckinsey/2012/12/03/big-data-advanced-analytics-success-stories-from-the-front-lines

espionage for adversarial reasons and unfair competition, signals intelligence collection and analysis[4] for (state) security reasons and customer privacy breach or exposure[5], either intended or accidental. This then forces to ensuring the security of the data and the privacy of the user, which we study next.

2.2 Data security

Securing data first aims at preserving its *integrity* and *confidentiality*, while not imposing constraints on its *availability* to authorised parties. Issues linked to its *authenticity* and *accountability* are related to its integrity, while data access is characterised by *non-repudiation* and *reliability*.

Data security is related to data usage and there, related services include *user authentication*, *user authorisation*, *access accountability* and *user reliability*. This finally mirrors to user privacy and trust, studied in detail in Sect. 4.1 and Sect. 4.2, respectively. The relationship between data security and user privacy is established via "guaranteeing privacy by securing access to private data". On a sociological level, the *Security Culture* (Alnatheer et al. 2012) is defined as the level of belief and expectations members of a group (e.g., an organisation or company) have regarding security. This is valid at all levels of the organisation and, in Ruighaver et al. (2007), it is demonstrated that the effectiveness of operational security policies in an organisation is positively correlated with the belief that the top management (decision makers) have in these policies. Greene and D'Arcy (2010) verify empirically the hypothesis that security culture (i.e., beliefs) and job satisfaction lead to a increased security behavior in the organisation.

A further distinction regarding the integration of data security is to be made between the public and private sector. The absence of economic markets for final product outputs in the public sector and the associated reliance on government financial resources place public agencies under strong political influence. As a result, information management systems in public organisations emphasise more the environmental factors rather than internal characteristics from the organisation. As demonstrated in Conklin (2007); Wang (2009), these differences play an important role in the diffusion of technology in e-government settings. In particular, decisions made on information security management in public organisations do not always follow technological rationales (Ruighaver et al. 2007).

Technologically, the challenge is to define security strategies in the *Information Management System* that will support business processes (see, e.g., Diesburg and Wang 2010 for technical surveys on digital data security). As given in Place and Hyslop (1982), Information management focuses on "plans and activities that need to be performed to control an organisation's records". Here, security should "ensure the continuity and minimise business damage by preventing and minimising the impact of security incidents" (Solms 2006, 2010). Authors of the latter references structure the evolution of security policies into several waves (illustrated in Fig. 2) showing the focus of every development phase, from purely technological to management-related concerns.

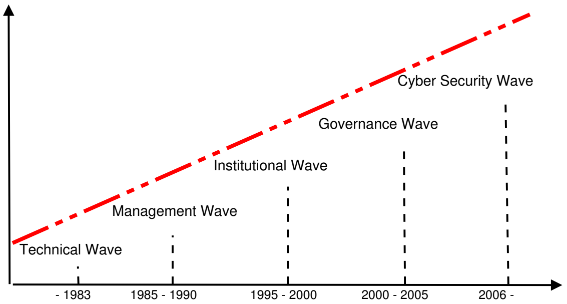

Figure 2: The 5 waves of Information Security (created after Solms 2010). The drift from technical to societal issues is clearly visible

From its origins (first wave), information security has been seen from a technical perspective. With

[4] e.g., the Echelon Network (2001) over communication networks, or the Prism Initiative (2013) over the Web

[5] e.g., the case of "AOL user 927" (2006)

the stability of the technical solutions, the question has moved onto the integration of security practices at a management (second wave), institutional (third wave) and governance (fourth wave) levels. The interweaving of private and public communication networks (e.g., private companies exposing themselves on the Web) has generated the related security issues of cybersecurity (fifth wave). Here, the construction of a secure context over highly complex interconnected communication networks (cybersecurity ERCIM News Special Theme: Cybercrime and Privacy Issues 2012) should go with the help of reference organisations such as the ISO/IEC (COBIT (Control Objectives for Information and related Technology) 4.1: Framework for IT Governance and Control Last retrieved: June 2013; ISO/IEC 27002:2005. Information technology – Security techniques – Code of practice for information security management 2005). These institutions supervise the creation of standards whose role is to protect an organisation's information asset in the context of confidentiality integrity and availability. Standards are generally largely biased, depending on economical, political or simply technical interests. In every domain, the debate of which standard is better always exists. Data security is no exception (see, e.g., Solms 2005). In the particular case of big data, valuable information is potentially hidden within massive amounts of data. Hence, in this context, defining the cost-benefit tradeoff of securing data is a hard task. As much as there are open technical challenges for securing data at very large-scale, there are also strategic open questions on the overall gain of such efforts. In the context of business informatics, the data often originates from customer behavior or input. Security questions therefore go beyond the technical benefits (e.g., the quality of data modelling), they encompass ethical issues, related to user rights issues related to data preservation (e.g., including privacy, as detailed in Sect. 4.1)

3 Data Intelligence

Business intelligence is related to an accurate use of the data collected at large scale. The main aspects of this adequate management is the accessibility and legibility of knowledge where available, and the creation of knowledge by automated or supervised processes. Figure 3 schematises the stages for the creation of knowledge from data, leading to accurate decision-making support.

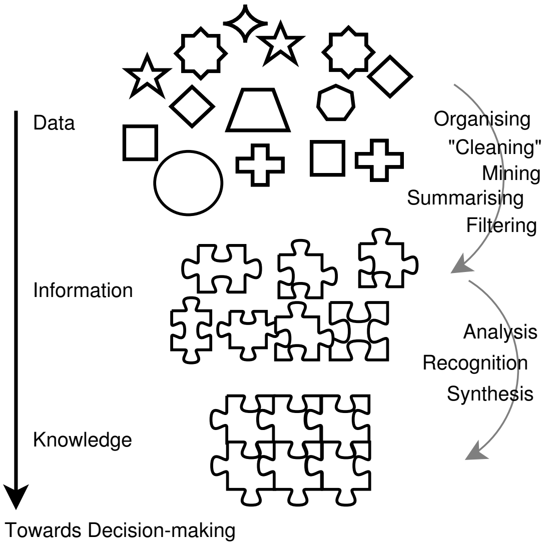

Figure 3: Creating knowledge from data

Data is the raw material that can be collected, either as characteristics (of users, products, *etc*) or as traces of activities (logs, sensing, results, *etc*). Information is obtained by compacting data into its coherent structures (e.g., patterns, summary). This step consists in aligning the data onto a model. Knowledge is obtained via processing Information with a high-level of understanding (e.g., semantic understanding). This step consists in matching the Information with known high-level concepts.

3.1 Knowledge Management

The domain of knowledge management, via the definition of extensive data description schemes offers solutions for accurate (semantic) query-based information or service access. The grail of *Knowledge Management Systems* (KMS) is reasoning and inference. From a non-redundant, but complete, knowledge representation, data content or actor behaviour may be predicted and linked together. However, the complexity and induced costs of design, creation, maintenance and compatibility of such descriptions generally impede their usage and development at large.

These strategies nevertheless find applications in well-understood, closed domains. Hence, besides their utility in representing knowledge via ontologies and inference models aligned with the semantic web, KMS (Abramowicz et al. 2010; Hu et al. 2010; Jiang et al. 2009) have been applied to specialised domains, including domains related to business. This is the case for enterprise modelling (Frank 2013) and cartography (Tribolet and Sousa 2013) and the modelling of business processes (Abramowicz et al. 2009; Sanz 2013). The reader is referred to the latter references for further insights in future developments of knowledge management in all aspects of business modelling.

3.2 Data mining and filtering

Data mining is the unsupervised (or weakly supervised, where weak assumption may be made over the inner structure of the data at hand) discovery of recurrent or coherent patterns in the data (Fayyad et al. 1996; Rajaraman and Ullman 2012). It develops in parallel to the field of *Machine Learning* (see, e.g., Domingos 2012) where the aim of the supervised process is to teach the machine specific decisions via the processing of examples. As such, data mining maybe coined as *Knowledge Discovery* (finding recurrent patterns in the data), complementing knowledge management, whereas machine learning is about *knowledge propagation* (extending known decisions to unknown situations), related to the field of *Predictive Analysis.* The flow of information evolves with time and trends develop. Such trends are characterised by the emergent surges of information patterns such as recurring keywords or phrases within text, or repeating events in usage logs. A particular case of data mining, suited to business informatics, is therefore *Emerging Trend Detection* (ETD) (Kontostathis et al. 2003). It studies flows of information along a timeline and extracts specific topic areas whose focus becomes more important at a point of time. It should be viewed as an automated mining process, since the manual inspection of flows of information at large-scale is simply not feasible. ETD is of crucial importance for data analysis, event prediction and decision-making in many areas, including business, finance, or politics. As such it is fully relevant for the analysis of big data. By identifying growing interests, actors of these domains will be able to react accordingly and even predict future evolution. For example, based on mining discussions over a social network, a company may decide to create a new product associated with a trendy product (e.g., sensitive pens for tablets) or, on the contrary, retract a product whose philosophy goes against current trends (e.g., large cars go against emerging "green" feeling). Investors may also anticipate fruitful niches if they can detect emerging trends at an early stage. They may also learn from the past by mining historical data to understand what caused the success or failure of such investment or product.

ETD generally considers documents as being aligned along a timeline and emerging trends also appearing and growing along that timeline (Ganesh et al. 2011). An early survey in (Kontostathis et al. 2003) lists and analyses systems proposed in the late 90's that operate on textual technical data such as the INSPEC database and the IBM DB2 US Patent database. In Le et al. (2005), a technique also applying on the scientific literature is proposed to track trendy topics using counts and bibliographic measures along time. Several temporal models have been proposed for the analysis of topics over time such as the *Dynamic Topic Model* (Blei and Lafferty 2006), *Topics* *over Time* (Wang and McCallum 2006) and the *Trend Analysis Model* (Kawamae 2011).

The medical literature, notably with the availability of the PubMed database is a domain of interest for ETD (Mörchen et al. 2008). Goorha and Ungar (2010) apply it to news wire articles, blogs posts, review sites and tweets, in search for interest rises in products or companies. A huge flow of information is processed daily based on word and phrases counts. Leskovec et al. (2009) correlate the appearance of given phrases in news with its occurrence in blogs. Similar studies have more recently be applied on Twitter data (e.g., Weng et al. 2010).

Collaborative and hybrid *recommender systems* (Park et al. 2012) leverage the wisdom of the crowd and propagate user interests across a community. They can result in the emergence or fall of an item, an idea by aggregating and propagating adequately consistent user judgements. Collaborative filtering operations may be seen as a local form of mining and trend detection within user interests. As such, they are also very close to the notion of crowdsourcing (see below). The main idea is to create a bipartite graph between products and customers where user ratings (judgments) are used as edge weights. Information is then propagated along this graph to group customers and/or products and thus, predict new edge weights (i.e., the judgement of a costumer over a product).

This framework for recommendation is used in Selke and Balke (2011) to cater for the lack of relevant or accurate information available to customers over "experience products". Authors demonstrate the effectiveness of their technique in the context of movie recommendations. This relates the idea of creating online and automatically item descriptions and therefore also relates to information retrieval. An early study on how such systems may be formally evaluated is proposed in Herlocker et al. (2004).

Whereas collaborative filtering uses implicit or explicit user judgements, sentiment analysis (a.k.a opinion mining explores blog texts, customer reviews or comments to track the acceptance or rejection of a product, an idea or a decision within a population (customers, voters, *etc*). Several approaches exist, including using *sentiment dictionaries* to map text words to opinions or sentiments with polarity (is/is not) (Liu 2012).

3.3 Information access: retrieval, filtering and browsing

Search and retrieval operations have installed themselves as a base paradigm for accessing items from within a repository. They are mostly based on the notion of a query formulation (Baeza-Yates and Ribeiro-Neto 2011). (Seidel et al. 2008), for example, demonstrates how such tools may support creativity in a business context.

Since precise data description is often a costly operation (or simply incompatible with the pace at which data is produced), in the case of systems operating over poorly described or non-textual data, the idea of *query-by-example* has emerged (Rui et al. 1998) as a help to construct accurate queries. Positive and negative examples are aggregated over intermediate search operations, in order to form a descriptive set for the sought items. Examples then become the base for online learning operations, so as to generalise *classes* of provided *relevant* and non-relevant items (Bruno et al. 2007; Wyl et al. 2011).

Browsing systems have been proposed and are also mostly based on the definition of a search objective (Heesch 2008). Such systems are typically oriented towards the localisation of a known information, be it media copy detection or user's *mental model* localisation (Ferecatu and Geman 2009) (see also Fig. 4). They iterate user *judgements* over appropriately-chosen sample sets of information to estimate the target item the user has in mind. This framework has been extended by (Lofi et al. 2010) from photo search to product browsing for mobile e-commerce.

Adaptive Hypermedia (AH) also relies on the navigation paradigm for information exploration to resolve the issue of complex query formulation. As accurately given in De Bra et al. (2004):

> "The core problem in finding the information you want, in all the above cases, is *describing* what you want. Results from search engines are often disappointing because most search requests are too short and unspecific to yield good results. Once a Web site with interesting information is found, it is often difficult to navigate to interesting pages only, because the site can only be navigated using its predefined link structure, independently of the search request that brought you to that site. The community of *user modelling* and *adaptive hypermedia* offers solutions for this problem: using information gathered about the user during the browsing process to change the *information content* and *link structure* on-the-fly. User modelling captures the mental state of the user, and thus allows that knowledge to be combined with the explicit queries (or links) in order to determine precisely what the user is looking for. To support the design of this user model-based adaptation, reference models like AHAM (De Bra et al. 1999; Wu 2002) and Munich (Koch and Wirsing 2002), both based on the Dexter Model by Halasz and Schwartz (1994), have been introduced in an attempt to standardise and unify the design of adaptive hypermedia applications, used mostly in isolated information spaces such as an online course, an electronic shopping site, an online museum, *etc*".

In Brusilovsky (2001), a taxonomy of AH technologies is further presented. The taxonomy is analysed in detail in Stash (2007), along with an extensive review of AH systems.

The above involves the notion of *user modelling* and a comprehensive review on personalisation research in e-commerce is presented in Adolphs and Winkelmann (2010).

Information filtering comes as a helper solution for the interactive formulation of search queries. Rules are defined over product characteristics, in order to define the class of the sought items as the intersection of solution sets for the rules. Rules are generally based on information *facets*. Facets are orthogonal, mutually exclusive dimensions of the data whose range is quantised in relevant intervals (Hearst 2008). Facets may be determined from the data model itself by highlighting important characteristics of the data. In exploratory conditions however, i.e., when the data is not fully understood, it may be interesting to let facets emerge automatically or interactively for providing interpretation of its organisation and to facilitate its exploration (Zwol and Sigurbjörnsson 2010). Several routes may be taken to automatically determine data facets. They all consist in using the data or a representative sample in a mining process to identify a reduced set of orthogonal projection operators whereby every data item is identified by its set of projections.

Faceted search is extensively used over e-commerce sites when products bear inherent orthogonal characteristics. For example, this is the case for real estate commerce with facets such as product type, surface, region, price range, ...

3.4 Visual analytics

The above tools are used to make sense of the data itself, using the intrinsic data content or the usage context of the data. *Visual analytics* refers to "the science of analytical reasoning facilitated by interactive visual interfaces" (Heer and Schneidermann 2012; Keim et al. 2010; Thomas and Cook 2006) and creates a link between content-based data mining and interactive data exploration techniques, as described above.

Visual Analytics supports the user in exploring the data and to interactively guide the system to find a formal solution that matches an intuitive solution (the mental model) to the problem (see Fig. 4).

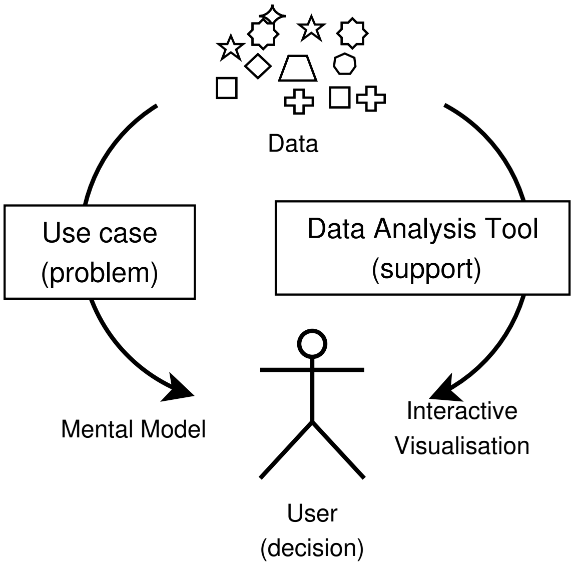

Figure 4: The process of Visual Analytics where the user is matching a mental model of the solution with the knowledge inferred from the data (see also Fig. 3)

In Zhang et al. (2012), a review of commercial systems for Visual Analytics, to support facing this big data era is proposed in the context of Business Intelligence. Various use cases are explored (inc. medical, microblogging) and performance over factors such as scalability and effectiveness for supporting decision-making are given. Authors then issue a number of future challenges related to effective large-scale data analysis.

One of the key parameters in interactive data analysis is to offer proper user interfaces and to adequately leverage the potential of user interaction, seen as a source of semantic knowledge into the system (Morrison et al. 2012). The role of users and user groups is studied in the next sections.

4 From the user to the community

While an accurate use of the data is fundamental to the decision process, the ultimate actor in the process remains the *user*. There are many user-related issues technologies should take care of. As much as data should be secured, the privacy of a user should be guaranteed. This will allow the user to act freely in the environment s-he is confronted with. Consequently, a consistent and reliable user behaviour may lead to trust and high reputation that may be used in many contexts for information access and recommendation.

Thanks to ever-developing communication media, users may also group into communities and form social groups. The emergence and identification of such social networks allows the analysis to move from the individual to the prototypical user-community s-he belongs to. This electronic crowd represents a task-force and a mass of semantic knowledge that crowdsourcing efforts aim at capturing.

4.1 User privacy

User privacy (Danezis and Gürses 2010; ERCIM News Special Theme: Cybercrime and Privacy Issues 2012; Hansen et al. 2008) is directly related to data security. The relationship between data security and user privacy is established via "guaranteeing privacy by securing access to private data".

User privacy can be understood as a two-fold concept, ethical and technological. Ethics should prevent the usage of user data to infer specific user needs and thus make that user fragile over communication networks. User data should be studied statistically and anonymously so that it returns to the user as a member of one user class, not as an individual. Many more ethical aspects should be defined in parallel of the advent of big data (Davis 2012). This is the role of governmental or not-for-profit independent organisations (e.g., the UN World Trade Organisation) to counter the temptation of inadequate usage of this data from large Internet companies, even though it is known that individuals value their privacy but tend to give it up easily as customers (Pogue 2011).

Technological solutions should ensure that the user data and behavior (e.g., mirrored into usage logs) remain private and are not accessible in their raw form to anyone. Anonymity may be a solution to privacy (Edman and Yener 2009), but again, this approach may not be feasible anymore, as soon as the individual becomes a customer or a user of social networks (De Cristofaro et al. 2012; Fung et al. 2010).

Also related to privacy is the possibility for secure authentication (Poller et al. 2012), preventing identity spoofing. These fields, associated to digital forensic and secured biometrics, directly relate to the notion of trust over communication networks.

4.2 Modelling trust

Trust is a social notion that an individual or a group (persons or organisation) develops over time and along experience. It measures the belief that the actions of an individual or a group may be predicted (e.g., from social knowledge of the individual or group) and stay within the limits of a predefined frame. Trust is closely related to the notion of reputation (Castelfranchi and Falcone 1998; Pinyol and Sabater-Mir 2013). It is opposed to the adverse behaviour of cheating via fraud and attacks (Hoffman et al. 2009). As such, the estimation of trust and reputation represents the estimation of a risk for the environment where the individual or group in question is active.

Models for trust and reputation over communication networks such as the Internet have been proposed with essentially two approaches. The *game theory approach* formalises a competition context where the objective is to maximise payoff with minimising risks. Trust estimation therefore relies on associated risk-minimisation tools. The *cognitive approach* accounts for elements such as beliefs, goals, desires and intentions. As such the resulting trust models bear as much value in their result than in their capability to explain the result. A thorough review of these models and their classification is proposed in Pinyol and Sabater-Mir (2013). These models are important to estimate the value of user interaction in systems such as recommender systems (Maida et al. 2012) (see also Sect. 3.3 above).

4.3 Social Network analysis

The constitution of social communities and groups of interest have allowed to move from the individual perspective to mass-address for business (essentially for push-based advertisement, the main revenue model for the Web). The study of social networks is therefore essential to structure the potential of such communities, including via the detection of key network features such as connectivity and influential nodes (Gomez-Rodriguez et al. 2012; Sun et al. 2013).

In relation to adaptive hypermedia and recommendation systems, where it is the study of user interaction that leads to recommendation, the study of social media (media hyperlinked in social networks) may allow the inference of recommendation (friends over Facebook or connections over LinkedIn) (Backstrom and Leskovec 2011). One of the difficulties here is the scale at which algorithms should operate. A compensating advantage of human-structured networks is their reputed low diameter (originally valued to 6 (Schnettler 2009), but said to be reduced to 4 over social networks) enabling local computations.

4.4 Social labor: Human Computation and crowdsourcing

There is a large labor potential to leverage over the Internet. This is known as *Human Computation* (Ahn 2005; Quinn and Bederson 2011) and also relates to crowdsourcing (Jones 2013). This strategy is, for example, used to help digitising characters via the ReCAPTCHA (Completely Automated Public Turing Test To Tell Computers and Humans Apart) system. Here, the trust in the user is evaluated by presenting a problem with a known answer. The answer to an unknown problem proposed simultaneously is then used as a statistical clue towards the right solution of this latter problem. In general, these tools, along with the *Games With a Purpose* (GWAP[6]), use the fact that human capabilities to perform (visual) pattern recognition surpass

[6] Games With a Purpose: http://www.gwap.com

by far that of an automated process, with the incentive of fun or commercial advantage. Recommender systems may also be seen as a form of crowdsourcing in that they seamlessly federate user judgements to create semantic information about items, products or services.

The impressive performance of such collaborative systems demonstrate the potential of labor to be federated over the internet. Another way of federating the crowd as a workforce is the use of digital labor (Larson et al. 2012). For example, the Amazon Mechanical Turk mediates between job requesters and workers. A requester creates a HIT (Human Intelligence Task) and proposes a reward for it. This HIT generally consists of a short but repetitive task such as asserting the presence of an object in an image. The trust into workers' competences may be evaluated by initial trials and a reputation system is active for both workers and requesters.

Eventually, if enough workers act on a simple task, this workforce constitutes a parallel processing machine (e.g., the *click-farms* to cheat Internet ads) and software APIs have even been developed to make that process fully transparent.

4.5 Social knowledge : Folksonomies

While human labor may be organised over the Internet, there are also several initiatives to federate human knowledge, following the "Wisdom of the Crowd" paradigm (Surowiecki 2004). Beyond the ever-growing Wikipedia and its collaborative edition model, including trust and reputation mechanisms, the combination of the semantic web and Web 2.0 for social behaviour enables the gathering of a social knowledge, known as *folksonomy* (Lohmann and Díaz 2012). This knowledge is made accessible accessible to machines via semantic web technologies and also offers a great potential for the development of adaptive or human-tailored business services.

5 Discussion and conclusion

Modern communication networks have fueled the creation of massive volumes of data. In this paper, we have discussed how this data may become an asset for business activities. This thorough overview of the information landscape, augmented with a large number of key references aims at providing a faithful picture and guideline for the practitioner who wants to attack the problem of data management and analysis in a business context. We highlighted and exemplified the potential benefits of data analysis but also the complexity and issues related to this task. We advocated for considering in parallel the data and the user viewpoints. Both perspectives share commonalities in their structure and analysis. The first being that most of the data originates from the users and that the users will then be modeled (in their behavior) via the analysis of data. Further, as much as data may be seen at different scales, user and user communities may be modeled at different scales. There is therefore much to gain in keeping this relationship alive when exploring and exploiting the data.

In this era of big data, large-scale data analysis becomes a strategic field of development. The promise of a *reasoning machine* by the field of artificial intelligence in the 1960's has been replaced by the statistical crunching of massive data with the side effect of smoothing out interesting details. The original three Vs of big data impose shallow processing for scalability. It is still an open challenge to design scalable process to filter (project or denoise, however a volume reduction strategy may be based on) the data to lower volumes and enable more effective analysis. In parallel, distributed infrastructures accommodating hierarchical processing of the data may help finding the essence of information and focus on these sparse *interesting needles* in the *data haystack*. It is also a commonplace that the potential of big data for business profitability is more an intuition than a frequent reality[7]. Hence,

[7] https://www.facebook.com/dan.ariely/posts/904383595868

not only effective processing and efficient infrastructures are needed but also the right analysis models are still lacking.

The ubiquity of communication platforms and networks have shaped the culture and raised new concerns and approaches towards privacy and security issues. Further deployments are likely to be guided by the further integration of connected hardware into our everyday lives. The news ways of interacting with the data that these devices create may be exploited for further augmenting the ubiquity and usefulness of the data for the customer. In turn, these powerful sensing devices will enable companies to tailor recommendation and targeting systems even further. Where the original Web was about simple data, the Web 2.0 about people and their relationships, the emergent *web of things* (suggested as “Web 3.0”) proposes to connect “Anything, Anywhere, Anytime”. Objects will be part of the communication process and their usage, location, proximity, *etc* will be tracked for better user behavior understanding. Extending the concept of *Tangible User Interface*, devices may symbol any data or virtual entity and be moved, exchanged, or combined as any object can be, with an associated impact in the virtual world.

Affective computing (Picard 2000), assessing user emotions via physiological sensors will also allow providers to penetrate even further into users' wishes. The trade-off between privacy and utility is thus very likely to continue evolving.

6 Acknowledgments

This work has been partly funded by Swiss National Science Foundation (SNSF) via the Swiss NCCR(IM)2 and the Swiss State Secretariat for Education and Research (SER) supporting ICT COST Action IC1002 “Multilingual and Multifaceted Interactive Information Access” (MUMIA) .

References

Abramowicz W., Haniewicz K., Kaczmarek M., Zyskowski D. (2009) Semantic Modelling of Collaborative Business Processes. In: Kusiak A., goo Lee S. (eds.) eKNOW. IEEE Comp. Soc., pp. 116–122

Abramowicz W., Fensel D., Frank U. (2010) Semantics and Web 2.0 Technologies to Support Business Process Management. In: Business & Information Systems Engineering 2(1), pp. 1–2

Adolphs C., Winkelmann A. (2010) A rigorous literature review on personalization research in e-commerce (2000-2008). In: Journal of Electronic Commerce Research 11(4), pp. 326–341

von Ahn L (2005) Human Computation. PhD thesis, Carnegie Mellon University Last Access: (UMI Order Number: AAI3205378)

Alnatheer M., Chan T., Nelson K. (2012) Understanding And Measuring Information Security Culture. In: Pacific Asia Conference on Information Systems (PACIS2012) Proceedings. Ho Chi Minh City, Vietnam

Backstrom L., Leskovec J. (2011) Supervised random walks: predicting and recommending links in social networks. In: Poceedings of the WSDM'2011 Conference, pp. 635–644

Baeza-Yates R., Ribeiro-Neto B. (2011) Modern Information Retrieval: the concepts and technology behind search, 2nd. Add. Wesley

Big Data: Science in the Petabyte Era. *Nature*, vol. 455, num. 7209. Last Access: (special issue)

Blei D. M., Lafferty J. D. (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. ICML '06. Pittsburgh, Pennsylvania, pp. 113–120

Bruno E., Kludas J., Marchand-Maillet S. (2007) Combining Multimodal Preferences for Multimedia Information Retrieval. In: Proceedings of the international workshop on Workshop on multimedia information retrieval

Brusilovsky P. (2001) Adaptive Hypermedia. In: User Modelling and User-Adapted Interaction 11, pp. 87–110

Castelfranchi C., Falcone R. (1998) Social Trust. In: Proceedings of the first workshop on deception, fraud and trust in agent societies. Minneapolis, USA

COBIT (Control Objectives for Information and related Technology) 4.1: Framework for IT Governance and Control Institute, Information Systems Audit and Control Association (ISACA) http://www.isaca.org

Conklin W. (2007) Barriers to Adoption of e-Government. In: System Sciences, 2007. HICSS 2007, pp. 98–98

Danezis G., Gürses S. (2010) A critical review of 10 years of privacy technology. In: Surveillance Cultures: A Global Surveillance Society? UK

Davis K. (2012) Ethics of Big Data. O'Reilly Media

De Bra P., Houben G.-J., Wu H. (1999) AHAM: A Dexter-based Reference Model for Adaptive Hypermedia. In: Proceedings of the 10th ACM conference on Hypertext and Hypermedia. Darmstadt

De Bra P., Aroyo L., Chepegin V. (2004) The Next Big Thing: Adaptive Web-Based Systems. In: Journal of Digital Information 5(1)

De Cristofaro E., Soriente C., Tsudik G., Williams A. (2012) Hummingbird: Privacy at the Time of Twitter. In: Security and Privacy (SP), 2012 IEEE Symposium on, pp. 285–299

Dean J., Ghemawat S. (2004) MapReduce: Simplified Data Processing on Large Clusters. In: OSDI'04: Sixth Symposium on Operating System Design and Implementation. San Francisco, CA

Diesburg S. M., Wang A.-I. A. (Dec. 2010) A survey of confidential data storage and deletion methods. In: ACM Comput. Surv. 43(1), 2:1–2:37

Domingos P. (Oct. 2012) A few useful things to know about machine learning. In: Commun. ACM 55(10), pp. 78–87

Edman M., Yener B. (Dec. 2009) On anonymity in an electronic society: A survey of anonymous communication systems. In: ACM Comput. Surv. 42(1), 5:1–5:35

ERCIM News Special Theme: Big Data. 89

ERCIM News Special Theme: Cybercrime and Privacy Issues. 90

Fayyad U., Piatetsky-Shapiro P., Smyth P. (1996) Knowledge Discovery and Data Mining: Towards a unified framework. In: Proceedings of the ACM SIG KDD Conference

Fayyad U. (2012) Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. In: Proceedings of the ACM SIGKDD Conference. (material available online – June 2013)

Ferecatu M., Geman D. (2009) A statistical framework for image category search from a mental picture. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 31(6), pp. 1087–1101

Frank U. (2013) Enterprise Modelling: the next steps. In: IEEE Conference on Business Informatics (IEEE–CBI 2013). Vienna, Austria

Fung B. C. M., Wang K., Chen R., Yu P. S. (June 2010) Privacy-preserving data publishing: A survey of recent developments. In: ACM Comput. Surv. 42(4), 14:1–14:53

Ganesh M. S., Reddy C. P., N.Manikandan, Venkata D. P. (2011) TDPA: Trend Detection and Predictive Analytics. In: International Journal on Computer Science and Engineering 3(3)

Gomez-Rodriguez M., Leskovec J., Krause A. (Feb. 2012) Inferring Networks of Diffusion and Influence. In: ACM Trans. Knowl. Discov. Data 5(4), 21:1–21:37

Goorha S., Ungar L. (2010) Discovery of significant emerging trends. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. Washington, DC, USA, pp. 57–64

Greene G., D'Arcy J. (2010) Security Culture and the Employee-Organization Relationship in IS Security Compliance. In: Proc. of the 5th Annual Symposium on Information Assurance. New York, USA

Halasz F., Schwartz M. (1994) The Dexter Hypertext Reference Model. In: Communications of the ACM 37(2), pp. 30–39

Hansen M., Schwartz A., Cooper A. (2008) Privacy and Identity Management. In: Security Privacy, IEEE 6(2), pp. 38–45

Hearst M. A. (2008) UIs for Faceted Navigation: Recent Advances and Remaining Open Problems. In: Workshop on Computer Interaction and Information Retrieval, HCIR. Redmond, WA

Heer J., Schneidermann B. (2012) Interactive Dynamics for Visual Analytics. In: Communication of the ACM 55(4)

Heesch D (2008) A survey of browsing models for content based image retrieval. In: Multimedia Tools and Applications 40 (2)

Herlocker J. L., Konstan J. A., Terveen L. G., Riedl J. T. (Jan. 2004) Evaluating collaborative filtering recommender systems. In: ACM Trans. Inf. Syst. 22(1), pp. 5–53

Hoffman K., Zage D., Nita-Rotaru C. (Dec. 2009) A survey of attack and defense techniques for reputation systems. In: ACM Computer Surveys 42(1), 1:1–1:31

Hu S., Wan L., Zeng R. (2010) Web2.0-based Enterprise Knowledge Management Model. In: Information Management, Innovation Management and Industrial Engineering (ICIII), 2010 International Conference on Vol. 4, pp. 476–480

ISO/IEC 27002:2005. Information technology – Security techniques – Code of practice for information security management (update of ISO/IEC 17799) International Standard Organisation

Jiang H., Liu C., Cui Z. (2009) Research on Knowledge Management System in Enterprise. In: Computational Intelligence and Software Engineering, CiSE 2009. International Conference on, pp. 1–4

Jones G. J. F. (2013) An introduction to crowdsourcing for language and multimedia technology research. In: Proceedings of the 2012 international conference on Information Retrieval Meets Information Visualization. PROMISE'12. Springer-Verlag, Zinal, Switzerland, pp. 132–154

Kawamae N. (2011) Trend analysis model: trend consists of temporal words, topics, and timestamps. In: Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, pp. 317–326

Keim D., Kohlhammer J., Ellis G., Mansmann F. (eds.) Mastering the Information Age – Solving Problems with Visual Analytics. Eurographic Digital Library

Koch N., Wirsing M. (2002) The Munich Reference Model for Adaptive Hypermedia Applications. In: 2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, pp. 213–222

Kontostathis A., Galitsky L., Roy S., Pottenger W. M., Phelps D. (2003) A survey of ETD in Textual Data Mining. In: Berry M. (ed.) A Comprehensive Survey of Text Mining. Springer

Larson M., Soleymani M., Eskevich M., Serdyukov P., Jones G. J. (2012) The Community and the Crowd: Developing large-scale data collections for multimedia benchmarking. In: IEEE Multimedia, Special Issue on Large-Scale Multimedia Data Collections

Le M.-H., Ho T.-B., Nakamori Y (2005) Detecting Emerging Trends from Scientific Corpora. In: International Journal of Knowledge and Systems Sciences 2(2)

Leskovec J., Backstrom L., Kleinberg J. (2009) Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD 2009. Paris, France, pp. 497–506

Liu B. (2012) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies 1 Vol. 5. Morgan & Claypool

Lofi C., Nieke C., Balke W.-T. (2010) Mobile Product Browsing Using Bayesian Retrieval. In: 12th Conference on Commerce and Enterprise Computing (CEC 2010). Shanghai, China, pp. 96–103

Lohmann S., Díaz P. (2012) Representing and visualizing folksonomies as graphs: a reference model. In: Proceedings of the International Working Conference on Advanced

Visual Interfaces (AVI'12). ACM, Capri Island, Italy, pp. 729–732

Maida M., Maier K., Obwegeser N., Stix V. (2012) A Multidimensional Model of Trust in Recommender Systems. In: EC-Web, pp. 212–219

Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A. H. (2011) Big data: The next frontier for innovation, competition, and productivity. Mc Kinsey Global Institute

Mohamed H., Marchand-Maillet S. (2012) Distributed media indexing based on MPI and MapReduce. In: Multimedia Tools and Applications, pp. 1–25

Mörchen F., Dejori M., Fradkin D., Etienne J., Wachmann B., Bundschus M. (2008) Anticipating annotations and emerging trends in biomedical literature. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. Las Vegas, Nevada, USA, pp. 954–962

Morrison D., Tsikrika T., Hollink V., de Vriesand E. Bruno A. P., Marchand-Maillet S. (2012) Topic modelling of clickthrough data in image search. In: Multimedia Tools and Applications, pp. 1–23

Mukherjee A., Liu B., Glance N. (2012) Spotting Fake Reviewer Groups in Consumer Reviews. In: Proceedings of the 21st International Conference on World Wide Web. WWW'12. Lyon, France, pp. 191–200

Park D. H., Kim H. K., Choi I. Y., Kim J. K. (2012) A Literature Review and Classification of Recommender Systems Research. In: Expert Systems with Applications (in press)

Picard R. W. (2000) Affective Computing. MIT Press

Pinyol I., Sabater-Mir J. (2013) Computational trust and reputation models for open multiagent systems: a review. In: Artificial Intelligence Review 40(1), pp. 1–25

Place I., Hyslop D. (1982) Records management: controlling business information. Reston Pub. Co.

Pogue D. (2011) Don't Worry about Who's watching. In: scientific American 304(1), p. 32

Poller A., Waldmann U., Vowe S., Turpe S. (2012) Electronic Identity Cards for User Authentication – Promise and Practice. In: Security Privacy, IEEE 10(1), pp. 46–54

Quinn A. J., Bederson B. B. (2011) Human computation: a survey and taxonomy of a growing field. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '11. ACM, Vancouver, BC, Canada, pp. 1403–1412

Rajaraman A., Ullman J. D. (2012) Mining Massive Datasets. Cambridge University Press

Rui Y., Huang T. S., Ortega M., Mehrotra S. (1998) Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval. In: IEEE Trans. on Circuits and Systems for Video Technology 8(5)

Ruighaver A., Maynard S., Chang S. (2007) Organisational security culture: Extending the end-user perspective. In: Computers and Security 26(1)

Sanz J. L. C. (2013) Enabling Customer Experience and Front-office Transformation through Business Process Engineering. In: IEEE Conference on Business Informatics (IEEE-CBI 2013). Vienna, Austria

Schnettler S. (2009) A structured overview of 50 years of small-world research. In: Social Networks 31(3), pp. 165 –178

Seidel S., Muller-Wienbergen F. M., Rosemann M., Becker J. (2008) A Conceptual Framework for Information Retrieval to Support Creativity in Business Processes. In: Proceedings 16th EuropeanConference on Information Systems. Galway, Ireland

Selke J., Balke W.-T. (2011) Turning Experience Products into Search Products: Making User Feedback Count. In: 13th IEEE Conf. on Commerce and Enterprise Computing (CEC 2011). Luxembourg

von Solms B. (2005) Information Security governance: COBIT or ISO 17799 or both? In: Computers and Security 24(2)

von Solms B. (2006) Information security – the Fourth Wave. In: Computer and Security 25(3), pp. 165–168 von Solms B. (2010) The 5 Waves of Information Security – From Kristian Beckman to the Present. In: Rannenberg K., Varadharajan V., Weber C. (eds.) Security and Privacy – Silver Linings in the Cloud. IFIP Advances in Information and Communication Technology Vol. 330. Springer Berlin Heidelberg, pp. 1–8

Stash N. (2007) Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System. PhD thesis, Eindhoven University of Technology, The Netherlands

Sun K., Morrison D., Bruno E., Marchand-Maillet S. (2013) Learning Representative Nodes in Social Networks. In: 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Gold Coast, AU

Surowiecki J. (2004) The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. New York: Doubleday

Thomas J., Cook K. (2006) A Visual Analytics Agenda. In: IEEE Computer Graphics and Applications 26(1), pp. 10–13

Tribolet J., Sousa P. (2013) Enterprise Governance and Cartography. In: IEEE Conf. on Business Informatics (IEEE-CBI 2013). Vienna

Vossen G. (2013) Big data as the new enabler in business and other intelligence. In: Vietnam Journal of Computer Science 1(1), pp. 1–12

Wang J.-F. (2009) E-government Security Management: Key Factors and Countermeasure. In: Information Assurance and Security, 2009. IAS09. Fifth International Conference on Vol. 2, pp. 483–486

Wang X., McCallum A. (2006) Topics over time: a non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD ’06. Philadelphia, PA, USA, pp. 424–433

Weng J., Lim E.-P., Jiang J., He Q. (2010) TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. Proceedings of the third ACM international conference on Web search and data mining WSDM ’10. New York, New York, USA, pp. 261–270

Wu H (2002) A Reference Architecture for Adaptive Hypermedia Applications". PhD thesis, Eindhoven University of Technology

von Wyl M., Mohamed H., Bruno E., Marchand-Maillet S. (2011) A parallel cross-modal search engine over large-scale multimedia collections with interactive relevance feedback. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval. ICMR ’11. ACM, Trento, Italy, 73:1–73:2

Zhang L., Stoffel A., Behrisch M., Mittelstadt S., Schreck T., Pompl R., Weber S., Last H., Keim D. (2012) Visual analytics for the Big Data era – A comparative review of state-of-the-art commercial systems. In: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, pp. 173–182

van Zwol R., Sigurbjörnsson B. (2010) Faceted exploration of image search results. In: Proceedings of the 19th international conference on World Wide Web (WWW’2010)

Stéphane Marchand-Maillet

Department of Computer Science
Centre Universitaire Informatique (CUI)
University of Geneva
Carouge
Switzerland
stephane.marchand-maillet@unige.ch

Birgit Hofreiter

Electronic Commerce Group
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna
Austria
birgit.hofreiter@tuwien.ac.at