# A Proposed Contextual Model for Big Data Analysis Using Advanced Analytics

**Manjula Ramannavar and Nandini S. Sidnal**

**Abstract** Big Data has numerous issues related to its primary defining characteristics of the three V's: Variety, Volume and Velocity. A greater segment of Big Data is attributed to semi-structured or unstructured text that emanates from social interactions on the web, emails, tweets, blogs, etc. Conventional approaches are overwhelmed by the data deluge and fall short to perform. These challenges consequently create scope for research in developing models to analyze data and extract actionable insights to realize the fourth V, i.e., Value. The purpose of this paper is to propose a contextual model for Resume Analytics that utilizes Semantic technologies and Analytic (Descriptive, Predictive and Prescriptive) procedures to find a befitting match between a job and candidate(s). The related work, issues and challenges and design requirements are presented along with a discussion of the analytical framework for the opted use case.

**Keywords** Big data · Descriptive analytics · Predictive analytics · Prescriptive analytics · Semantic technologies

## 1 Introduction

The Internet, web, social media, smartphones and other online work based-tools generate humongous digital data affecting our daily lives. Big Data refers to massive datasets, ranging beyond terabytes, which are heterogeneous and generated at an ever-increasing rate. Unstructured text contributes to around eighty per cent [13] of the total digital data generated by companies due to machine generated logs, call detail records, e-commerce, emails, etc. The conventional tools and technologies are beleaguered by the distinctive dimensions of Big Data [16]. Hence efficient

M. Ramannavar (✉) · N.S. Sidnal
KLS Gogte Institute of Technology, Belagavi, India



V.B. Aggarwal et al. (eds.), *Big Data Analytics*, Advances in Intelligent Systems and Computing 654, https://doi.org/10.1007/978-981-10-6620-7_32

algorithms, methods and tools need to be devised to capture, process and analyze this data in an intelligent automated fashion to draw inferences. Advanced Analytics [15] entails application of a multitude of analytic processes to address the diversity of Big Data to yield responsive information in the form of descriptions, predictions and prescriptions. The inferences drawn may guide strategic decision-making to enable an organization to gain a competitive edge. Big Data and Advanced Analytics thus come together to become critical enablers of the modern economy.

## *1.1 The Resume Analytics Problem*

Human Resources (HR) [25] are an important asset to any organization. The strength of an organization lies in its right HR workforce. A Resume or Curriculum Vitae represents the functional aspects of a job aspirant and consists of several sections, namely qualification, experience, skills, personal details, etc. It is used to screen applicants; the ones that get through are selected for subsequent round, generally, an interview. In the context of automated Resume Analytics, the following two stakeholders are identified:

**Job Provider**: This role is represented by an individual, a group or an organization. A Job Provider is often faced with the problem of finding the talent pool, i.e., the most befitting candidate to fill up a vacancy. A job description indicates the requirements, i.e., qualification, skill set and experience needed for a job. In the event of a vacancy, job provider calls for posts online by floating job descriptions.

**Job Seeker**: This role is represented by an individual who is a job aspirant. A prospective candidate prepares a resume to reflect his personality in an appealing manner and uploads it, anticipating a call for the recruitment process.

The Resume Analytics problem is to find the most appropriate fit between a job description and prospective candidate(s), given a pool of resumes and a job description. Commercial tools such as ALEX [1], Daxtra [9], RChilli [20], Sovren [24], Textkernel_hr_suite [27] are available for job and resume parsing and mapping a job to resume. However, open-source contributions to the specified problem are solicited for the benefit of the masses. This proposition is an attempt towards developing a suite of components to deal with the Resume Analytics problem.

The rest of the paper is organized as follows: Sect. 2 reviews related work. Section 3 identifies issues and challenges based on literature survey. The design requirements for the model are enumerated in Sect. 4. Section 5 utilizes the Big Data Value Chain or the Big Data Analysis Pipeline to suggest the intended model, and Sect. 6 concludes the paper with directions for future work.

## 2 Related Work

This section reviews the literature to summarize the contributions made by various researchers towards resume parsing and allied fields.

The authors in [2] have worked on ontology-based ranking of documents using Graph Databases. Attribute values are extracted from resumes and mapped onto RDF. Resumes are ranked based on cosine similarity measures. Experiments for retrieval time for three cases are compared: Single RDF, Four RDFs and Three RDFs.

Latent Semantic Analysis (LSA) and ontological concepts are used to support e-recruitment in the teaching domain in [4]. Ontology is built, and vector is generated by applying LSA. Semantic similarity is found between job posting and CV.

Ontology based Resume Parser (ORP) system is proposed in [7] for Kariyer.net company that has more than 6,000,000 unstructured and freestyle résumés as MS Word. The architecture, working mechanism, similarity of concept, matching techniques and inference mechanism are introduced, and a case study for a Turkish resume is presented. The proposed ORP system in [7] is implemented in [6] which transforms an input résumé into structured format by splitting it into explicit segments, parsing it, normalizing it and finally applying classification and clustering task. [8] extends work done in [6] by incorporating Semantic Web Rule Language (SWRL) inference mechanism.

Machine learning-based CV Parser system is used in [10] to extract information from Hungarian, English and German CVs for a Career Portal. The system converts CVs into text files and extracts relevant data using maximum-entropy Markov model (MEMM).

Deductive model and an Ontology-Based Hybrid Approach are used to match job seeker and a posting in [11]. Similarity-based approach is used to rank applicants.

Most of the methods extract information from resume while in [17], a resume is selected from a group of resumes by extracting Special Skills type and Special Skill Values thereby improving performance.

Linked Data allows resources to be openly published and well connected to other resources thereby enabling discovery of highly relevant information. The authors in [18] publish Resume data in RDF into Web of Data using the Linked Data approach. Information may be consumed by machines using RDF and by humans using HTML. Issues of heterogeneity, interoperability and data reusing between multiple data sources are addressed using Linked Data.

A Semantic Web enabled System for Résumé Composition and Publication is implemented in [19]. The system assists a user to write a resume using domain ontologies and also annotate with reference to a common ontological vocabulary. It also enables a user to create his own tag-bag which would better expose the candidate on the web for crawlers.

A Framework is proposed in [21] for semantic annotation of Urdu web documents based on domain ontology. It deals with Urdu web documents that are free format, imprecise and in unintelligible formats.

In [22], a system for Candidate-Task matching in the E-Recruitment field is detailed. The system uses a web-based interface that allows one to enter required skills and desired skills and retrieve candidate(s) accordingly.

A system for information extraction from a resume based on Named Entity Clustering algorithm is presented in [23]. A resume is segmented. Chunkers recognize named entities which are then clustered and normalized.

Large Scale Skill Matching through Knowledge Compilation is the core theme in [28]. Knowledge Compilation approach is used where the knowledge base in Description Logics is converted into relational database. SQL queries are then executed over the database for skill matching. Job request is semantically matched against the candidates. Strict requirements and preferences may be specified.

It can be seen that most of the works utilize domain ontology to characterize background knowledge for information extraction and semantic representation. The focus is more on transforming an unstructured resume into a structured semantic format to facilitate information exchange and expert finding on the web. The authors in [13] discuss a host of machine learning techniques and methods used for automated categorization of text documents. According to editorial [14], Linked Data and Big Data together constitute the 4th paradigm in computing. Linked Data is a part of the Big Data Landscape and an ideal test bed for research on Big Data issues. There is a necessity to adopt the Big Data approach to solve the use case.

## 3 Issues and Challenges

A Big Data initiative needs to have a well-planned strategy and the right set of tools to cope with heterogeneity, speed and scalability concerns. A study of related work in Sect. 2 indicates the following research issues and challenges for the Resume Analytics problem:

- Resumes are uploaded in different formats at various sites. Collection and aggregation of these resumes is a challenge.
- A resume is free text and hence is of semi-structured or unstructured nature. It typifies the Variety dimension of Big Data. Suitable methods need to be devised to efficiently store, analyze and extract knowledge.
- The concepts of a resume and underlying relationships between them need to be clearly understood. This issue may be handled by capturing domain knowledge through ontologies.
- The information extracted from a resume has to be stored to facilitate querying and retrieval. Resource Description Framework (RDF) and Linked Data technologies can be used to enable data from different sources to be connected and queried.

- Resumes and job descriptions are continually uploaded by the two stakeholders all over the world. The Volume and Velocity aspects call for solutions to deal with the issues of scale and speed. Scalable-distributed computing frameworks have to be exploited to address these issues.
- Finding the most appropriate fit between job description and resume(s) necessitates innovative ways of applying techniques of Advanced Analytics in order to perform computational analyses in an optimal and effective manner.

## 4 Design Requirements

In the light of the above-stated issues, existing studies in allied disciplines have raised the need for a model [26] to aid the stakeholders in the recruitment process. Considering these domain-driven motivational needs, the following design (functional and non-functional) requirements ought to be met:

### *4.1 Functional Requirements*

**To articulate formalized repository of domain knowledge**: A job description and a resume embed domain knowledge of distinctive aspects such as objective/summary, skill set, experience, education. Distributed architectures and domain-based ontology need to be exploited to allow optimal storage and retrieval.
**To identify relevant concepts/entities related to the recruitment process**: To extract different sections and build a concept map. Synonyms, for example qualification and education and same terms expressed in different forms have to be identified as the same concept.
**To perform Resume Analytics**: To discover correlations (explicit and hidden) between extracted concepts based on the concept map. For example, Hierarchical relationship: Skill → Programming Languages → C. Ranks need to be assigned on the basis of qualitative evaluation of resumes using techniques of Advanced Analytics. Job description also needs to be ranked. Resumes are to be clustered according to the ranks.
**To present the extracted knowledge**: To present the results in the form of resumes satisfying the job requirements along with implications. Also seek the most appropriate fit between a job and candidate(s).

### *4.2 Non-functional Requirements*

**Scalability, Reliability and Availability**: It is needed to develop solution(s) that operate in parallel and distributed architectures offering fault-tolerance and replication.
**Resilience and Speed**: The solution(s) should provide quick responses and provide an acceptable level of service in the event of faults.

## 5 Proposed Model for Big Data Analysis

The proposed model in Fig. 1 meets the aforementioned functional requirements while the architecture in Fig. 2 satisfies the non-functional requirements of the above-stated design requirements. The Big Data Value Chain or the Big Data Analysis Pipeline [3] serves as the blueprint for the proposed model for Big Data Analytics. It consists of multiple distinct phases: Acquisition/Recording,

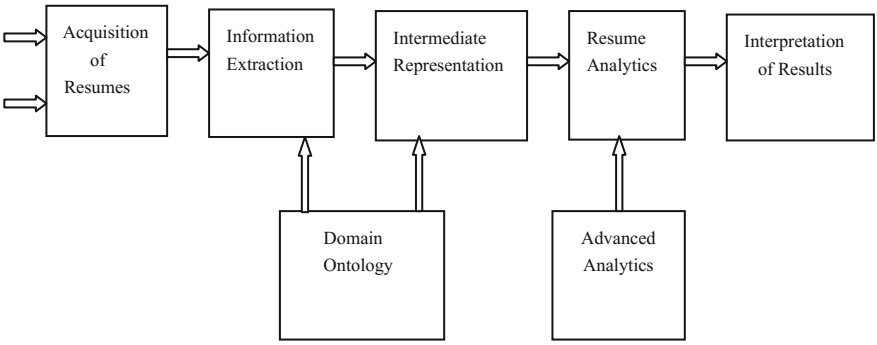
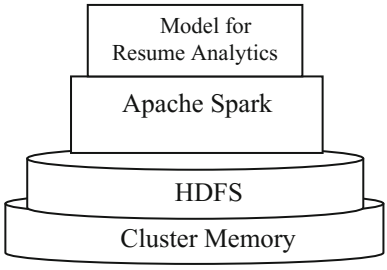



**Fig. 1** Proposed model for Big Data Analysis

**Fig. 2** Architecture of the proposed model



Information Extraction and Cleaning, Data Integration, Aggregation, and Representation, Query Processing, Data modelling, and Analysis and Interpretation. Figure 1 gives a picture of the proposed model for the Resume Analytics problem.

## 5.1 Phases of the Model

**Acquisition**: Resumes are collected from various sources such as emails, job portals, LinkedIn and recorded for the subsequent phases.
**Information Extraction**: Relevant concepts pertaining to skills, experience, qualification etc. are extracted from the resumes to build a concept map. Background knowledge in the form of domain ontology assists in semantic extraction of information.
**Intermediate Representation**: The extracted information is stored in an intermediate representation suitable for subsequent analysis. The representation should be chosen to enable quick and easy access. RDF and Linked Data technologies may be used for optimal semantic storage and retrieval.
**Resume Analytics**: Every resume in the intermediate representation is analyzed using the techniques of Advanced Analytics based on the various criteria against the job requirement and assigned a rank.
**Interpretation of Results**: The resumes satisfying the job requirements are visualized as results along with decision enabling indicators. The most appropriate fit between the job description and the candidate(s) also serve as the analytic results.

## 5.2 Analytics

There are three types of analytics: Descriptive analytics is a set of processes and technologies that summarize data to infer what is happening now or what has happened in the past. Predictive analytics is a set of processes that look into the past to predict the future. Prescriptive analytics not only enables to look into the future but also suggests actions to benefit from the predictions and shows the decision maker the implications of each decision option. While descriptive analytics look into the past, Advanced Analytics in the form of predictive and prescriptive analytics provide a forward-looking perspective and make use of techniques encompassing a wide range of disciplines including statistics, machine learning, optimization, simulation. The proposed work intends to integrate Advanced Analytics models to map a job to resume(s).

Figure 2 depicts the architecture to be used for the Resume Analytics Model.

Apache Spark [5] is an open-source scalable cluster computing platform that is much faster than Hadoop [16] due to in-memory computing primitives. It allows data to be directly loaded into cluster memory. Spark is also well suited to machine learning algorithms and can interface with a variety of distributed storage including Cassandra, Hadoop Distributed File System (HDFS), and Amazon S3. HDFS [12] is a highly available, highly fault-tolerant distributed file system designed to operate on commodity hardware. The proposed model shown in Fig. 1 would be implemented on top of Spark.

## 6 Experimental Results

In this section, the results of preliminary work are presented. To evaluate the performance, the work was applied on real data set of resumes, containing around 200 resumes of postgraduate and undergraduate students. Experiment is based on resume information where a dataset containing description of resume information is generated, and then, search queries are used to find the required information from the resume. This dataset is assumed as a part of the Web of Data and is used for experimental purpose. A CV dataset contains information about business and academic information, skill, company, title, etc.

Resume processing like building concepts and analyzing phase is done in Hadoop. To build concepts of resumes, multiple resumes are processed through MapReduce jobs. The qualitative measure 'coverage' can be taken as value denoting how many sections and subsections are covered in a resume. The next job was then to see how many concepts from a class are covered in a resume. This required extraction of concepts and performing matching. If all the concepts are covered in the resume, as per resume grammar, then that particular resume is considered as complete resume or else incomplete. To compute comprehensibility, each line in resume is fed to mapper which gives comprehensibility of that line. Reduce sums up all the results and gives comprehensibility of each resume as output. Figure 3 shows the selected resumes that match a job description.

Once MapReduce jobs are done, it gives summary of each resume called as Comprehensive Quality. It gives summary on missing information of header and data. After summing up the comprehensibility, each resume is ranked according to the coverage and comprehensibility. Finally, for a given job description, it lists all the resumes along with the ranks as shown in Fig. 4.

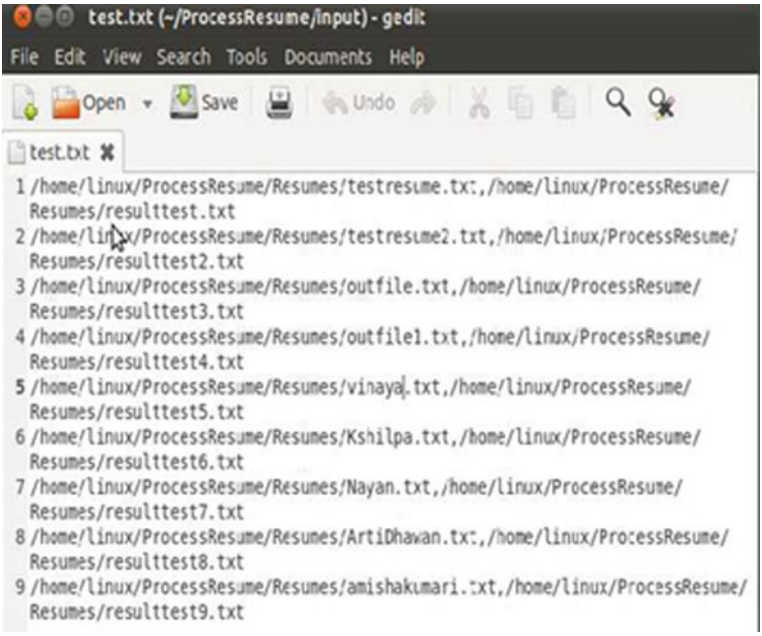



**Fig. 3** Selected resumes that match job description

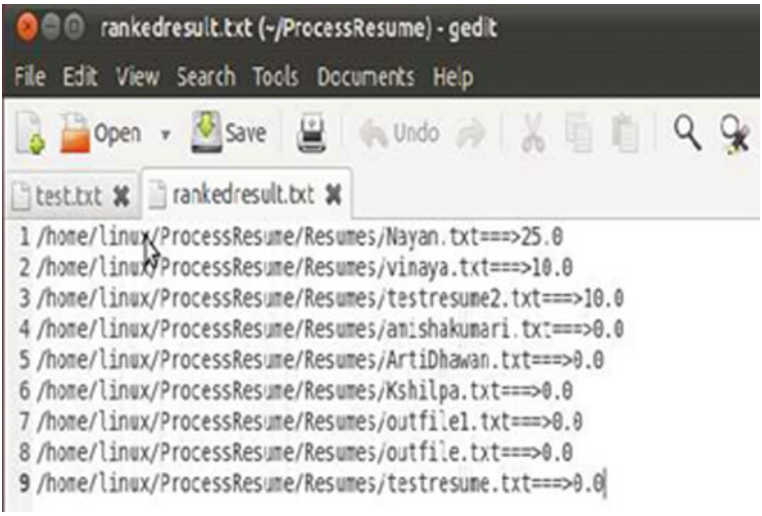



**Fig. 4** Ranked resumes

## 7 Conclusion and Future Work

Big Data and Analytics are the front runners of innovation and technology advancements. While Big Data is inevitable, analytic processes probe through it (which otherwise remained unexplored) to infer insights of unprecedented worth. Semi-structured or unstructured data contributes to a major chunk of Big Data. Managing it is a huge challenge that implores several avenues for research. This paper considered the case study of Resume Analytics and reviewed various contributions in its realm. A research proposition has been formulated stating the objectives and design requirements. The planned architecture and model to realize the objectives were also discussed.

Future work would focus on a deeper study of the constituent elements of the model. Attempts would be made towards developing a prototype for the various modules incorporated in the model from implementation perspective. The model would also have to be evaluated taking into account associated metrics followed by rigorous validations.

## References

1. ALEX Resume Parser: http://www.hireability.com/ALEX
2. Abirami, A.M., Askarunisa, A., Sangeetha, R.: Ontology based ranking of documents using graph databases: a big data approach, smarter planet and big data analytics workshop. In: Co-located with International Conference on Distributed Computing and Networking, 4 Jan 2014, Amrita University, Coimbatore
3. Agrawal, D., et. al.: Challenges and opportunities with big data. A community white paper developed by leading researchers across the United States through Collaborative Writing between Nov 2011 to Feb 2012
4. Al-lasassmeh, K.Q., Kayed, A.K.A.: Latent Semantic analysis (LSA) and ontological concepts to support e-recruitment. Department of Computer Science, Faculty of Information Technology, Middle East University, June 2013
5. Apache Spark: https://spark.apache.org/
6. Çelikelik, D., Elçi, A.: An ontology-based information extraction approach for Résumés. In: ICPCA-SWS 2012, LNCS 7719, pp. 165–179 (2013). Springer, Berlin
7. Çelikelik, D.: Towards a semantic based information extraction system for matching résumés to job openings. Computer Engineering Department, Istanbul Aydin University, Turkey
8. Çelikelik, D., et. al.: Towards an information extraction system based on ontology to match Résumés and jobs. In: 2013 IEEE 37th annual computer software and applications conference workshops, Kyoto, Japan, 22–26 July 2013. doi:10.1109/COMPSACW.2013.60
9. Daxtra Intelligent Recruitment Solutions: http://www.daxtra.com
10. Farkas, R., Dobó, A., Kurai, Z., Miklós, I., Nagy, A., Vincze, V., Zsibrita, J.: Information extraction from Hungarian, English and German CVs for a career portal. In: Prasath, R., et al. (eds.) MIKE 2014, LNAI 8891, pp. 333–341 (2014). Springer International Publishing, Switzerland
11. Fazel-Zarandi, M., Fox, M.S.: Semantic matchmaking for job recruitment: an ontology-based hybrid approach. In: 8th International Semantic Web Conference (2009)
12. Hadoop Distributed File System (HDFS): https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

13. Halper, F., Kaufman, M., Kirsh, D.: Text analytics: the Hurwitz victory index report. Technical Report (2013). http://www.sas.com/news/analysts/Hurwitz_Victory_Index-TextAnalytics_SAS.PDF
14. Hitzler, P., Janowicz, K.: Linked data, big data, and the 4th Paradigm. Editorial. Semantic Web Journal by IOS Press. Feb 2013. http://www.semantic-webjournal.net/system/files/swj488.pdf
15. Kaisler, S.H., Espinosa, J.A., Armour, F., Money, W.H.: Advanced analytics—issues and challenges in the global environment. In: 47th Hawaii international conference on system science, Hilton Waikoloa, Big Island, pp. 729–738, 6–9 Jan 2014. doi:10.1109/HICSS.2014.98
16. Katal, A.,Wazid, M., Goudar, R.H.: Big data: issues, challenges, tools and good practices. In: Sixth International IEEE Conference on Contemporary Computing (IC3), Noida, India, pp. 349–353, 8–10 Aug 2013
17. Maheshwari, S., Sainani, A., Krishna Reddy, P.: An approach to extract special skills to improve the performance of resume selection. In: 6th International Workshop on Databases in Networked Information Systems (DNIS2010), Centre for Data Engineering, International Institute of Information Technology, Hyderabad, India, Mar 2010
18. Marjit, U., Sharma, K., Biswas, U.: Discovering resume information using linked data. Int. J. Web Semant. Technol. (IJWesT) **3**(2) (2012)
19. Mirizzi, R., Noia, T.D., Sciascio, E.D., Michelantonio, T.: A Semantic Web enabled System for Résumé Composition and Publication. In: 3rd IEEE International Conference on Semantic Computing (ICSC 2009), Berkeley, CA, USA, 14–16 Sept 2009. doi:10.1109/ICSC.2009.40. http://www.researchgate.net/publication/221406004_A_Semantic_Web_Enabled_System_for_Rsum_Composition_and_Publication
20. Overview of RChilli Resume Parser Copyright © 2014 RChilli Inc.: http://rchilli.com/wp-content/uploads/2014/12/Overview-of-RChilli-Resume-Parser.pdf
21. Rajput, Q.: Ontology based semantic annotation of Urdu language web documents. In: 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2014), Gdynia, Poland, 15–17 Sept 2014, pp. 662–670. doi:10.1016/j.2014
22. Saellstrom, P.: A system for candidate-task matching in the e-recruitment field. Umea University, Department of Computing Science, Sweden, July 2013
23. Sonar, S., Bankar, B.: Resume parsing with named entity clustering algorithm. http://www.slideshare.net/swapnilmsonar/resume-parsing-with-named-entity-clustering-algorithm
24. Sovren Resume/CV Parser: http://www.sovren.com
25. Suen, H.: The effect of end user computing competence on human resource job performance: mapping for human resource roles. Afr. J. Bus. Manage. **6**(28), 8287–8295 (2012). doi:10.5897/AJBM11.577. ISSN 1993-8233 © 2012 Academic Journals. http://www.academicjournals.org/AJBM
26. Tao, J., Deokar, A.V., El-Gayar, O.F.: An Ontology-based information extraction (OBIE) framework for analyzing initial public offering (IPO) prospectus, pp. 769–778. In: 47th Hawaii International Conference on System Science, Waikoloa, HI, USA, 6–9 Jan 2014. doi:10.1109/HICSS.2014.103
27. Textkernel_hr_suite. Empowering Recruitment: www.textkernel.com
28. Tinelli, E., Colucci, S., Giannini, S., Sciascio, E.D., Donini, F.M.: Large Scale Skill Matching Through Knowledge Compilation. Foundations of Intelligent Systems, Lecture Notes in Computer Science, vol. 7661, pp. 192–201 (2012). http://link.springer.com/chapter/10.1007%2F978-3-642-34624-8_23