Information Processing and Management 48 (2012) 1124–1135

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A hybrid approach to managing job offers and candidates

Rémy Kessler [a,*], Nicolas Béchet [c], Mathieu Roche [d], Juan-Manuel Torres-Moreno [b], Marc El-Bèze [a]

[a] LIA/Université d'Avignon et des Pays de Vaucluse, 339 chemin des Meinajariès, 84911 Avignon, France
[b] École Polytechnique de Montréal, CP 6079, succ. Centre-ville, Montréal (Québec) Canada H3C 3A7
[c] INRIA Domaine de Voluceau, BP 105, 78153 Le Chesnay Cedex, France
[d] LIRMM, CNRS Université Montpellier 2, 161 rue Ada, 34392 Montpellier, France

ARTICLE INFO

Article history:
Received 9 February 2011
Received in revised form 1 March 2012
Accepted 6 March 2012
Available online 10 April 2012

Keywords:
Natural language processing
Automatic summarization
Information retrieval
Human resources
Statistical approaches
Similarity measures

ABSTRACT

The evolution of the job market has resulted in traditional methods of recruitment becoming insufficient. As it is now necessary to handle volumes of information (mostly in the form of free text) that are impossible to process manually, an analysis and assisted categorization are essential to address this issue. In this paper, we present a combination of the E-Gen and CORTEX systems. E-Gen aims to perform analysis and categorization of job offers together with the responses given by the candidates. E-Gen system strategy is based on vectorial and probabilistic models to solve the problem of profiling applications according to a specific job offer. CORTEX is a statistical automatic summarization system. In this work, E-Gen uses Cortex as a powerful filter to eliminate irrelevant information contained in candidate answers. Our main objective is to develop a system to assist a recruitment consultant and the results obtained by the proposed combination surpass those of E-Gen in standalone mode on this task.

1. Introduction

The evolution of the job market has resulted in that traditional methods of recruitment becoming insufficient. The Internet has introduced a new way of managing human resources. Theoretically, shifting job search and recruitment activities to the Internet improves the quality of job matching by reducing search costs, increasing contact opportunities and rationalizing the screening process of job applicants (Marchal, Mellet, & Rieucau, 2007). Over the last few years, there has been a significant expansion of online recruitment (e.g. August 2003: 177,000 job offers, May 2008: 500,000 job offers).[1] The Internet has become essential in this process because it allows a better flow of information, either through job search sites or by e-mail exchanges. Nowadays, job seekers can send their curriculum vitae (CV) directly to companies (by e-mail or uploaded to dedicated servers on the Web). The job search task is becoming easier and less time consuming. The Internet makes every user a potential job seeker. Employees may be constantly in search of new career opportunities and job candidates may provide more interaction than can be managed efficiently by companies (Bourse, LeclFre, Morin, & Trichet, 2004). As intellectual capital has become one of the most strategic assets of successful organizations in the last decade, the capability of managing people's expertise, skills and experience represents a key factor in facing up to the increasing competitiveness of the global market (Colucci et al., 2003). Even though a browser has become a universal and easy tool for users, they

* Corresponding author.
E-mail addresses: remy.kessler@univ-avignon.fr (R. Kessler), nicolas.bechet@inria.fr (N. Béchet), mathieu.roche@lirmm.fr (M. Roche), juan-manuel.torres@univ-avignon.fr (J.-M. Torres-Moreno), marc.elbeze@univ-avignon.fr (M. El-Bèze).

[1] http://www.keljob.com.

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2012.03.002

frequently have to enter data into Web forms from paper sources and the need to "copy and paste" data between different applications is symptomatic of the issues of data integration. In this context, electronic recruitment tends to automate matching between the published information about the candidates and job offers. The *Laboratoire Informatique d'Avignon* (LIA),[2] the *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier*(LIRMM),[3] and Aktor Interactive[4] are developing the E-Gen system to resolve this issue. E-Gen is a Natural Language Processing (NLP) and Information Retrieval (IR) system composed of three main modules:

1. The first one extracts the information from a corpus of e-mails of job offers from Aktor's database.
2. The second module analyses the candidate's answers (i.e. splitting e-mails into cover letter (CL) and curriculum vitae).
3. The third module analyses and computes a relevant ranking of the candidate's answers.

Our first work (Kessler, Torres-Moreno, & El-Bèze, 2007) presented the first module: the identification of different parts of a job offer and the extraction of relevant information (type of contract, salary, localization, etc.). The second module analyses the content of a candidate's e-mail, using a combination of rules and machine learning methods (Support Vector Machines, SVM) and was presented in Kessler, Torres-Moreno, and El-Bèze (2008b). Furthermore, it separates the distinct parts of CV and CL with a precision of 0.98 and a recall of 0.96. Reading a large number of candidate answers for a job is a very time consuming task for a recruiting consultant. In order to facilitate this task, we propose a system capable of providing an initial evaluation of candidate answers according to various criteria. We do not seek the best or even a good candidate as no scoring is involved, but simply a candidate who has a close application to those already selected. Our previous work (Kessler, Béchet, Roche, El-Bèze, & Torres-Moreno, 2009) presented an approach based on a process of relevance feedback, permitting a reinforcement learning (Sutton & Barto, 1998). In this paper, we present an original combination of the E-Gen and CORTEX systems. Each document contains a number of additional information, present in many applications and which is partially removed by classical pre-processing. Each application added by the process of relevance feedback adds relevant information but also multiplies additional information. CORTEX allows us to filter these sentences and keep only the most relevant sentences at the evaluation step. Some related studies are briefly discussed in Section 2. Section 3 shows a general system overview. In Section 4, we describe the E-Gen pre-processing task, the strategy used to rank the candidate answers with relevance feedback and the coupling of E-Gen with the CORTEX summarization system. In Section 5, we present statistics about the textual corpus, experimental protocol, an example of CL summary generated by CORTEX, and several results.

2. Related work

Many approaches have been proposed in the literature to reduce the costly and tedious task of managing human resources. Candidate answers to a job offer come as *ad hoc* documents, and require semantic approaches to analyse them. The *BONOM* system is based on an indexing method (Morin, LeclFre, & Trichet, 2004; Cazalens & Lamarre, 2001). This method consists in using distributional attributes of documents to locate each part for the final indexation of the document.

A semantic-based method to select candidate answers and to discuss the economic impacts on the German government was proposed by Tolksdorf, Mocho, Heese, Oldakowski, and Christian (2006). In the same way (Gorenak & Mlaker KaF, 2010), perform a comparison between Slovenian, German, and British online job advertisements (ads). More recently (Marchal et al., 2007), present a comparison between French and English job search sites and newspapers as well as the various shortcoming of current matching systems. They propose a comparative analysis of job offers posted on the Internet with those posted in newspapers and they observe that search engine toolkits have a considerable impact on ad content which is generally more standardized and quantified than before.

Mocho, Paslaru, and Simperl (2006) discuss the relevance of a common ontology (HR ontology) to work efficiently with this kind of document. Using the same model (Dorn & Naz, 2007), outline a HR-XML based prototype dedicated to the job search task. The prototype selects and favors relevant information (paycheck, topic, abilities, etc.) from many job-service websites, such as `Jobs.net`, `aftercollege.com`, `Directjobs.com`, etc. Bourse et al. (2004) describe an efficient model and a management tool used for the selection of candidate-answers. They propose a prototype job portal which uses semantically annotated job offers and applicants to obtain a more accurate job search with query approximation.

The limitations of current systems for automatic selection of candidate answers are presented in Rafter, Bradley, and Smyt (2000). They propose a system based on collaborative filters (*ACF*) to automatically select profiles of candidate answers on the JobFinder website. Enrica and Iezzi (2006) present a model for ranking skills in the field of information technology in Italy with multidimensional scaling and cluster analysis. In the same way, Colucci et al. (2003) present a semantic based approach to the issue of skills detection in an ontology supported framework. Based on Description Logics formalization and reasoning, they propose a skill matching approach with contradiction matches and partial matches between skill profiles. Loth et al.

[2] http://www.lia.univ-avignon.fr.
[3] http://www.lirmm.fr.
[4] A French recruitment agency specialized in recruiting on the internet, (http://www.aktor.fr).

(2010) combine, through the SIRE project (Semantics-Internet-Recruitment-Employment) a linguistic approach and machine learning methods to perform an extraction of key terms of job ads in order to improve the categorization of each job offer.

The study of the most relevant document – the CV – to use it automatically has been a major subject of research. Ben Abdessalem Karaa (2009) presents a system for analyzing and structuring CVs with an extension of General Architecture of Text Engineering (GATE[5]). They obtain good results in precision/recall for each part of the document (personal information, experience, skill, and so forth) on a small corpus of CVs in French. Yahiaoui, Boufaïda, and Prié (2006) provide a semantic approach to generating some annotations of CVs and job offers with the help of a specialized ontology to match graduates and the level of a job offer. They present interesting results on a sample of data. Clech and Zighed (2003) propose a data mining approach. Their aim is to build automats which recognize CV topologies and candidate/job offer profiles. A first step differentiates the CV of employed executives from other CV. They use a specific term extraction to obtain a categorization with the C4.5 decision tree algorithm (Quilan, 1993). This method focuses on the specificity of selected terms or concepts, such as education level or relevant abilities, to build a classifier. The results of this method are still poor (an accuracy between 0.5–0.6 of correctly categorized CV). Roche and Kodratoff (2006) and Roche and Prince (2008) have made a terminology study of corpus composed of CVs (of the Vediorbis company (http://www.vediorbis.com)). Their approach extracts collocations from a CV corpus based on syntactic patterns such as Noun-Noun, Adjective-Noun, etc. Then, these collocations are ranked according to relevance to build a specialized ontology.

There are few studies on the treatment of the cover letter. Audras and Ganascia (2006) use cover letters to detect the usual errors in the field of acquisition of written French as a foreign language. The approach proposed is the detection of syntactic patterns particular to a group of learners, and which are absent or little used among native speakers. The study focuses in part on cover letter writing. Among the innovative solutions on the market, Twitter[6] has launched the job search site http://www.twitterjobsearch.com based on the concept of short messages (less than 140 characters) and ZaPoint[7] with an original solution, SkillsMapper, which transforms each CV into graphic format with various curves (training, education, etc.). In this paper, we present an approach to the application ranking by using a combination of similarity measures, relevance feedback and summaries of a CV and CL. Our approach is distinguished from other work by a purely statistical approach as well as reinforcement learning through the process of relevance feedback.

3. System overview

Nowadays technology proposes new approaches to the online employment market. E-Gen is a system which meets this challenge as fast and judiciously as possible. We chose emails as the input format, which is the most frequent mode of communication in this field. An e-mail inbox receives messages sometimes with an attached file containing the job offer. When a job offer is published online, a particular segmentation is required by the job search sites. Firstly, the job offer language is identified by using *n*-grams. Then, E-Gen parses the e-mail, splits the job offer into thematic segments, and retrieves relevant information (contract, salary, starting date, location, etc.) to generate an XML document for the job offer. Subsequently, a filtering and lemmatisation process is applied to the text, and is represented in a vector space model (VSM). A categorization of text segments (preamble, skills or profile, mission) is obtained by using a SVM classifier (Fan, Chen, & Lin, 2005). This preliminary classification is then transmitted to a "corrective" post-process which improves the quality of the solution (Module 1, described in Kessler et al., 2007). Preliminary experiments showed that segment categorization without segment position in job posting is not enough and may be a source of errors. In order to avoid this kind of error, we have decided to consider each job posting as produced by a succession of states in a Markov machine and we have applied a post-processing, based on the Viterbi algorithm (Viterbi, 1967). During the publication of a job offer, Aktor generates a temporary e-mail address for applying to the job. Each e-mail is redirected to human resources software (Gestmax[8]) to be read by a recruiting consultant. At this step, E-Gen analyses the candidate's answers to identify each part of the application and extracts the text from the e-mail and attached files (by using wvWare[9] and pdftotext[10]).

After a pre-processing task, we use a combination of rules and machine learning methods to separate each distinct part (CV or CL). We use a vector representation of each document with a label (CV or CL). With a learning set of 2.000 documents of each type, the system gets very good performance (F-score between 0.95 and 0.98). This process (Module 2 represented by the lowest box in Fig. 1) is more fully described in Kessler et al. (2008b). Once the CL and CV have been identified, the CORTEX system is applied to each document (Cover Letter and CV) and a summary is generated by concatenating high-scoring sentences. Afterwards, E-Gen performs an automated profiling of this application by using measures of similarity and a small number of applications that have been previously validated as relevant by a recruitment consultant (Module 3). The whole chain is summarized in Fig. 1.

[5] http://gate.ac.uk/.
[6] http://twitter.com.
[7] http://www.zapoint.com.
[8] http://www.gestmax.fr.
[9] http://wvware.sourceforge.net.
[10] http://www.bluem.net/downloads/pdftotext_en.

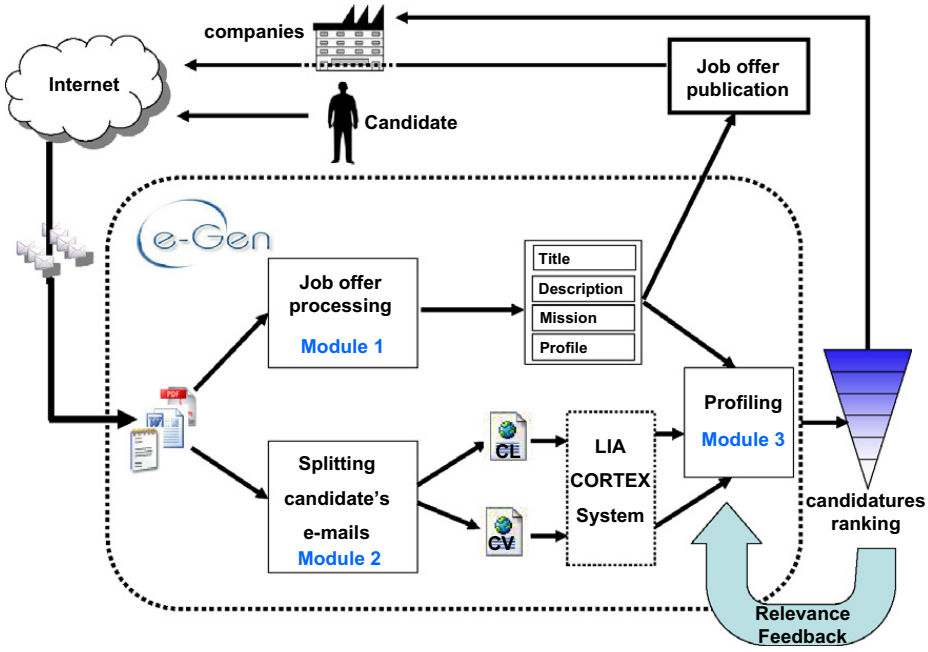

Fig. 1. System overview.

4. Coupling E-Gen profiling module and the CORTEX system

4.1. E-Gen profiling module

4.1.1. Linguistic pre-processing

Firstly, we remove information such as e-mail adresses, the names of candidates, addresses, names of cities in order to ensure that the applications become anonymous. Then, classic pre-processing is applied to textual information (job offer, CV, and CL). French accents are deleted and capital letters are converted to lower case. This pre-processing task is performed to obtain a representation well suited for the Vector Space Model (VSM). In order to avoid the introduction of noise into the models, the following items are also deleted: verbs and functional words (to be, to have, to need, etc.), common expressions with a stop word[11] list (for example, that is, each of, etc.), numbers (in numeric and/or textual format), symbols such as "$", "#", "∗". Finally, lemmatisation[12] is performed to significantly reduce the size of the lexicon. All these processes allow us to represent the collection of documents through the bag-of-words paradigm (a matrix of frequencies of terms (columns) for each candidate answer (rows)). To improve filtering, we tried parsing applications with different significant terms (like "Personal Information", "Education", "Work Experience", etc.) and extract only paragraphs with the relevant information, but initial tests showed a decline in results due to the great variability of signifiant terms and order of paragraphs.

4.1.2. Proximity between applications and job offer using similarity measures

After the step of linguistic pre-processing, each document is transformed into a vector with weights characterizing the frequency of terms *Tf*. Some tests with *Tf-idf* (Salton & Mcgill, 1986) were made but they offered no improvement. We have established a strategy using measures of similarity, to rank all applications in relation to a job offer. We combined different similarity measures between the candidate's answers (CV and CL) and the associated job offer. We decided to use several similarity measures as defined in Bernstein, Kaufmann, Kiefer, and Bnrki (2005): Cosine (Eq. (1)), which calculates the angle between job offer and each candidate answer, Minkowski distances (Eq. (2)) ($p = 1$ for Manhattan, $p = 2$ for Euclidean). The last measure used is Okabis (Eq. (3)) (Bellot & El-Bèze, 2001). Based on the formula of Okapi (Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1994), this measure is often used in Information Retrieval. To combine these measures, we use an Algorithm Decision (AD) (Boudin & Torres Moreno, 2007), which weights the values obtained by each measure of similarity. Several other similarity measures (Overlap, Enertex, Needleman-Wunsch, Jaro-Winkler, Jensen-Shannon divergence) have been tested but they are not retained in this study, because the results obtained were disapointing. All measures used and their combinations are described in Kessler, Béchet, Roche, El-Bèze, and Torres-Moreno (2008a).

[11] http://sites.univ-provence.fr/veronis/donnees/index.html.

[12] Lemmatisation finds the root of verbs and transforms plural and/or feminine words into masculine singular form. So we conflate terms *developer, development, developing, to develop* into develop.

$$\text{cosine}(j,d) = \frac{\sum_{i=1}^{n} j_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} j_i^2 \cdot \sum_{i=1}^{n} d_i^2}} \tag{1}$$

$$\text{Minkowski}(j,d) = \frac{1}{1 + \left(\sum_{i=1}^{n} |j_i - d_i|^p\right)^{\frac{1}{p}}} \tag{2}$$

$$\text{Okabis}(j,d) = \sum_{i \in d \cap j} \frac{\sum_{i=1}^{n} j_i \cdot d_i}{\sum_{i=1}^{n} j_i \cdot d_i + \frac{\sqrt{|d|}}{M_d}} \tag{3}$$

where j is a job offer, d is a candidate answer, i a term, j_i and d_i occurrence of i respectively in j and d, and M_d their average size.

4.1.3. Relevance Feedback

We previously changed the system to incorporate a process of Relevance Feedback (Sparck Jones, 1970). Relevance Feedback is a standard method used particulary for manual query reformulation. For example, the user carefully checks the answer set resulting from an initial query, and then reformulates the query. Rocchio's algorithm (Rocchio, 1971) and variations have found wide usage in Information Retrieval and related areas such as Text Categorisation (Joachims, 1997). Relevance Feedback has been proposed in Smyth and Bradley (2003) to help the user to find a job with server logs from the jobFinder site.[13] In our system, Relevance Feedback takes into account the recruiting consultant's choice during a first evaluation of a few CVs. Our goal is not a system capable of finding the best candidate, but a system capable of reproducing the judgement of the recruitment consultant. It is critical for recruiters not to miss a promising candidate that they may have unfortunately rejected. The goal of this Relevance Feedback approach is to help them to avoid this kind of error. We assume that successful candidates have similar profiles or, at least, that they have much in common. This approach uses documents returned in response to a first request to improve the search results (Salton & Buckley, 1990). In this case, we randomly take a few candidate answers (1–6 in our experiments) from all relevant candidate answers. These selected candidate answers are added to the job offer. So, we use manual Relevance Feedback to reflect user judgements in the resulting ranking. We increase the vector representation with the terms from the candidates considered relevant by a recruitment consultant. The system will recompute the similarity between the candidate's answer that we evaluate and the job offer enriched with relevant candidates. This allows Sim' to be recalculed for each measure of similarity between the application evaluated and the job offer expanded by **relevant** applications of the relevance feedback process:

$$Sim'_{\text{measure}}(j,d) = Sim_{\text{measure}}(j, d \| p_1 \| \cdots \| p_n) \tag{4}$$

where j is a job offer, d is a candidate's response, p_i is a **relevant** candidate's response, n are numbers of retained applications for Relevance Feedback and $\|$ is the concatenation operator.

The results, presented in Kessler et al. (2009) and hereafter called *ISMIS Result* showed an improvement in the quality of the ranking obtained for each application added to the process of relevance feedback. However, we suspected that a lot of unnecessary information was still kept in the evaluation and we wanted to use a filter to take into account the content of sentences. Each document contains additional information (hobbies, greeting and complimentary close, etc.) and standard pre-processing only partially removes it. The idea was to use a system of automatic summarization, coupled to E-Gen, as a powerful filter capable of removing non-essential information contained in CV and Cover Letters.

4.2. The CORTEX summarization system

Automatic summarization is useful to cope with ever increasing volumes of information. An abstract is, by far, the most concrete and recognized kind of text condensation. However, the CV is already a kind of summary, with a very important structure. We suspect that the filtering system of automatic summarization may not be useful in this case. Since the CL is in free text, we used CORTEX (Torres-Moreno, St-Onge, Gagnon, El-Bèze, & Bellot, 2009, 2001), an efficient state-of-art summarization system, in order to retain the more informative segments of the CL.

Each document of the application is transmitted to the CORTEX system which provides a summary based on the requested size. CORTEX is a document extract summarization system using an optimal decision algorithm that combines several metrics. These metrics result from processing statistical and informational algorithms on the document vector space representation. Fig. 2 presents an overview of the system.

The idea is to represent the text in an appropriate vectorial space and apply numeric processings to it. In order to reduce complexity, a pre-processing of the document is performed: words are filtered, lemmatized, and stemmed. Based on the terms that remain in the text after filtering, a frequency matrix γ is built in the following way: Each element γ_i^{μ} of this matrix represents the number of occurrences of the word i in the sentence μ.

[13] JobFinder (http://www.jobfinder.com).

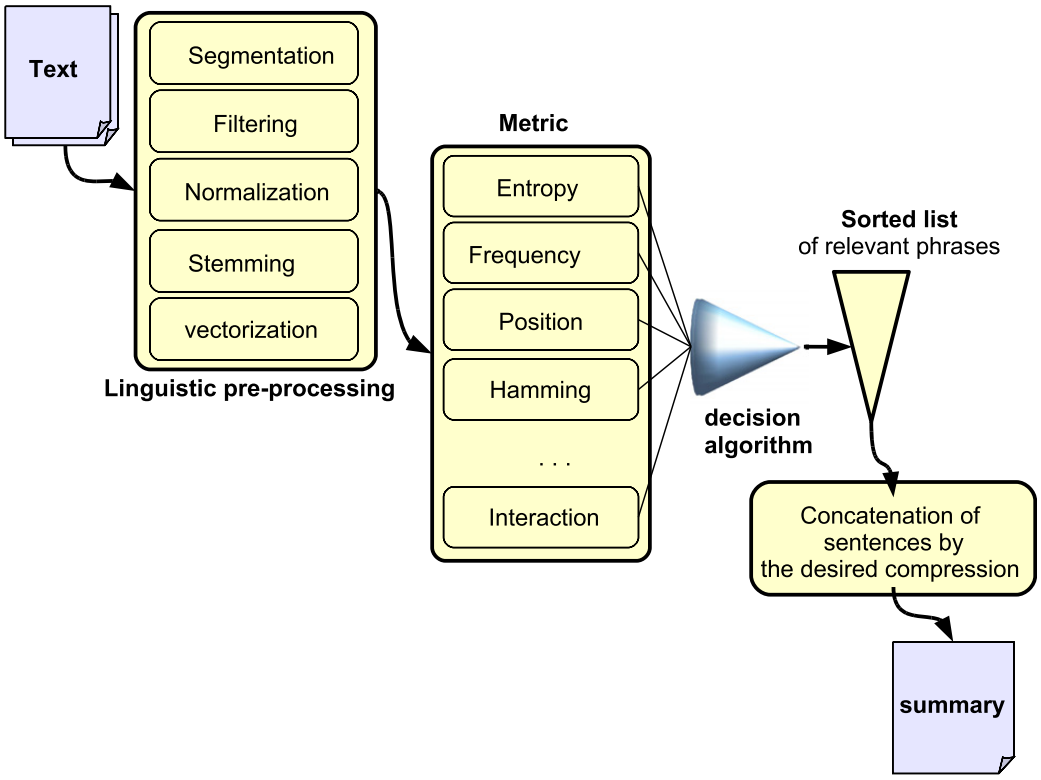

Fig. 2. CORTEX overview.

$$\gamma = \begin{bmatrix} \gamma_1^1 & \gamma_2^1 & \cdots & \gamma_i^1 & \cdots & \gamma_{N_L}^1 \\ \gamma_1^2 & \gamma_2^2 & \cdots & \gamma_i^2 & \cdots & \gamma_{N_L}^2 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \gamma_1^\mu & \gamma_2^\mu & \cdots & \gamma_i^\mu & \cdots & \gamma_{N_L}^\mu \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \gamma_1^{N_S} & \gamma_2^{N_S} & \cdots & \gamma_i^{N_S} & \cdots & \gamma_{N_L}^{N_S} \end{bmatrix}, \quad \gamma_i^\mu \in \{0, 1, 2, \ldots\} \tag{5}$$

Another matrix ξ, called *a binary virtual* or *presence matrix*, is defined as:

$$\xi_i^\mu = \begin{Bmatrix} 1 & \text{if } \gamma_i^\mu \neq 0 \\ 0 & \text{elsewhere} \end{Bmatrix} \tag{6}$$

Each line of these matrices represents a sentence of the text. Matrices γ and γ^T are the frequency matrix of the sentences and frequency matrix of the titles respectively.

The CORTEX system can use up to $\Gamma = 11$ metrics (Torres-Moreno, Velazquez-Morales, & Meunier, 2002) to evaluate the sentence's relevance.

The system scores each sentence with a decision algorithm which relies on the normalized metrics. Two averages are calculated, a positive $\lambda_s > 0.5$, and a negative $\lambda_s < 0.5$ tendency (the case $\lambda_s = 0.5$ is ignored). The following algorithm combines the vote of each metric:

$$\sum^s \alpha = \sum_{v=1}^{\Gamma} (\|\lambda_s^v\| - 0.5); \quad \|\lambda_s^v\| > 0.5$$

$$\sum^s \beta = \sum_{v=1}^{\Gamma} (0.5 - \|\lambda_s^v\|); \quad \|\lambda_s^v\| < 0.5$$

Γ is the number of metrics and v is the index of the metrics. The value given to each sentence s is calculated with:

$$if \left(\sum^s \alpha > \sum^s \beta \right)$$

then $Score_s^{cortex} = 0.5 + \sum^s \alpha / \Gamma$: retain s
else $Score_s^{cortex} = 0.5 - \sum^s \beta / \Gamma$: not retain s

The sentences are then ranked according to the obtained values. Depending on the desired compression rate, the sorted sentences will be used to produce the summary. The CORTEX system is applied to each document (Cover Letter) and a summary is generated by concatenating high-scoring sentences. We generated several abstracts with a variable compression rate (5%, 10%, 20%, ..., 50%, 75% of the size of the documents, in sentences) in order to test the impact of our powerful filter on the E-Gen system. The entire process chain is illustrated in Fig. 1. The best compression rates are generally with 30% (Torres-Moreno et al., 2009). The results are presented in Section 5.3.

5. Experiments

We selected a data subset from Aktor's database composed of 1917 candidates. This subset is called the *Mission Corpus*. It has a size of 10 MB of raw texts and contains 1,375,000 words. The *Mission Corpus* is composed of a set of 12 job offers covering various themes (jobs in accountancy, business, computer science, etc.) and their candidates. Each Job Offer is associated with at least six candidates identified as **relevant**. As described in Kessler et al. (2008a), each document is segmented to keep the relevant parts (we remove the description of the company (D) for the job offer). Each candidate answer is tagged as **relevant** or **irrelevant**. A **relevant** value corresponds to a potential candidate for a specific job chosen by the recruiting consultant. An **irrelevant** value is associated with an unsuitable candidate for the job (this is a decision made by the manager of a human resources company). Our study was conducted on French job offers because the French market represents Aktor's main activity. Table 1 shows a few statistics about the *Mission Corpus*.

5.1. Example of CL summaries

Fig. 3 presents[14] an example of an original Cover Letter and Fig. 4. Its corresponding summary[15] generated by the CORTEX system with a 30% compression rate (in number of sentences).

All the documents of *Mission Corpus* were previously made anonymous. We observe that the original CL contains a number of useless information for ranking, such as addresses, phone numbers or form of address at the beginning or end of the letter. The last part of the CL is generally as "Yours faithfully", "Yours sincerely", "Best regards", all of which represent irrelevant information. We further observe in Fig. 4 that the summary obtained with CORTEX removes all this information.

5.2. Experimental protocol

We measured the similarity between a job offer and its candidate's responses. These measures (Section 4.1.2) rank the candidate's answers by computing a similarity between a job offer and the associated candidate answers. We use the ROC curves to evaluate the quality of the ranking obtained. ROC curves (Ferri, Flach, & Hernandez-Orallo, 2002) come from the field of signal processing. They are used in medicine to evaluate the validity of diagnostic tests. In our case, ROC curves show the rate of irrelevant candidate answers on the X-axis and the rate of relevant candidate answers on the Y-axis. The

[14] Pierre ASPRE

26 years old
19 Verdun street 92870 Vannes
06-06-06-06-06.
Subject: collaboration offer
Vannes, November 27th, 2008
Dear Sir,
The Accountant is a key player not only for the proper functioning of the enterprise, but also in increasing profitability. With his legal knowledge in tax and social issues, he can make substantial savings: he is a key player for maintaining a cash reserve by ensuring the payment of customer invoices and knowing how to deal with the late settlement of invoices.
Therefore I offer my skills. They allow me to:
– Manage with rigueur the accounts of a company.
– Ensure legal compliance activities (payroll, tax billing etc.).
– Provide advice particularly important in times of assessment, all thanks to my seriousness, my strength and my analysis.
I suggest we meet to discuss all the terms of our future cooperation.
I look forward to hearing from you.
Best regards.
Pierre ASPRE.

[15] Pierre ASPRE

Subject: collaboration offer
The Accountant is a key player not only for the proper functioning of the enterprise, but also in increasing profitability. With his legal knowledge in tax and social issues, he can make substantial savings: he is a key player for maintaining a cash reserve by ensuring the payment of customer invoices and knowing how to deal with late settlement of invoices.
– ensure legal compliance activities (payroll, tax billing etc.).
– provide advice particularly important in times of assessment, all thanks to my seriousness, my strength and my analysis.

Table 1
Mission corpus statistics.

Number	Job title	Number of candidate answers	Number of	
			Relevant	**Irrelevant**
34861	Sales engineer	40	14	26
31702	Accountant, department suppliers	55	23	32
33633	Sales engineer	65	18	47
34865	Accountant assistant	67	10	57
34783	Accountant assistant	108	9	99
33746	3 chefs	116	60	56
33553	Trade commissioner	117	17	100
33725	Urban sales consultant	118	43	75
31022	Recruitment assistant	221	28	193
31274	Accountant assistant junior	224	26	198
34119	Sales assistant	257	10	247
31767	Accountant assistant junior	437	51	386
Total		1917	323	1594

Pierre ASPRE
26 ans
19 Avenue Verdun 92870 Vannes
06-06-06-06-06.
Objet : offre de collaboration.
Vannes, le 27/11/2005
Monsieur,
Le comptable est un acteur essentiel non seulement au bon fonctionnement de l'entreprise, mais aussi dans l'accroissement de la rentabilité. En effet, grâce à ces connaissances juridiques en matière fiscale et sociale, il permet de réaliser des économies substantielles: il est un des acteurs principaux du maintien d'une réserve de trésorerie en assurant le paiement des factures clients et en sachant jouer sur les délais de réglement des factures fournisseurs.
C'est pourquoi je vous propose mes compétences. Elles me permettent de :
- gérer de manière rigoureuse les comptes d'une entreprise.
- veiller à la conformité légale des actions (paie, fiscalité, facturation....
- prodiguer des conseils particulièrement importants en période de bilan, le tout grâce à mon sérieux, mon dynamisme et mon analyse.
Je vous propose de nous rencontrer afin de discuter ensemble des modalités de notre future collaboration.
Dans cette attente, je vous prie de recevoir l'expression de mes salutations distinguées.
Pierre ASPRE

Fig. 3. Example of full Cover Letter.

Pierre ASPRE
Objet: offre de collaboration.
Monsieur, Le comptable est un acteur essentiel non seulement au bon fonctionnement de l'entreprise, mais aussi dans l'accroissement de la rentabilité. En effet, grâce à ces connaissances juridiques en matière fiscale et sociale, il permet de réaliser des économies substantielles: il est un des acteurs principaux du maintien d'une réserve de trésorerie en assurant le paiement des factures clients et en sachant jouer sur les délais de réglement des factures fournisseurs.
- veiller à la conformité légale des actions (paie, fiscalité, facturation.
- prodiguer des conseils particulièrement importants en période de bilan, le tout grâce à mon sérieux, mon dynamisme et mon analyse.

Fig. 4. Summary of Cover Letter (see Fig. 3) at a 30% compression rate.

Area Under the Curve (AUC) can be interpreted as the effectiveness of a measurement of interest. In the case of candidate answers ranking, a perfect ROC curve corresponds to obtaining all relevant candidate answers at the beginning of the list and all irrelevant ones at the end. This situation corresponds to AUC = 1. The diagonal line corresponds to the performance of a random system, progress of the rate of relevant candidates being accompanied by an equivalent degradation in the rate of irrelevant candidates. This situation corresponds to AUC = 0.5, as explained in Fawcett (2006). An effective measurement

Table 2
Results of CL or CV according to the compression rate of Cortex and part of job offer (with or without Description part).

CORTEX compression rate (%)	CV + DTMP	CV + TMP	CL + DTMP	CL + TMP
100 (full text)	**0.622**	**0.648**	**0.567**	**0.560**
75	0.565	0.575	0.563	0.556
50	0.558	0.569	0.553	0.560
40	0.552	0.565	0.561	0.565
30	0.549	0.560	**0.569**	**0.571**
20	0.520	0.558	0.564	0.566
10	0.559	0.559	0.543	0.554
5	0.550	0.542	0.521	0.523

Table 3
Results for CV and cover letter according to the compression rate.

CORTEX compression rate (%)	CV and CL summaries		Full CV and CL summary	
	DTMP	TMP	DTMP	TMP
100 (full text)	**0.634**	**0.642**	0.634	0.642
75	0.521	0.581	0.639	0.641
50	0.556	0.551	0.643	0.649
40	0.544	0.568	0.643	0.651
30	**0.570**	**0.587**	**0.646**	**0.653**
20	0.569	0.533	0.641	0.652
10	0.564	0.534	0.631	0.645
5	0.546	0.547	0.638	0.649

of interest to order candidate's answers consists in obtaining the highest AUC value. This is strictly equivalent to minimizing the sum of the ranks of the relevant candidate's answers. ROC curves are resistant to imbalance (for example, an imbalance in the number of positive and negative examples) (Roche & Kodratoff, 2006). For each job offer, we evaluated the quality of the ranking obtained by this method. Candidate answers considered are only those composed of CV and CL.

5.3. Results

In this section, we present the results obtained by combining the CORTEX system with the E-Gen ranking application. CORTEX was used as an additional filter which generates a summary of each document before E-Gen evaluation. We keep the structure of data for job offers as described in Kessler et al. (2008a). A job offer is composed of a Description (D), a Title (T), a Mission (M), and a Profile (P). For these experiments, we use two combinations of a job offer content, keeping only Title, Mission, Profile (TMP) and all information of a job offer (DTMP). Results are presented in Tables 2 and 3. Each column presents a part of the application with different sizes of summaries for each line (75%, 50%, ..., 5%). Full text is a result obtained with 100% of the document and was published previously in Kessler et al. (2008a, 2009).

Table 2 presents results obtained for each part of the application separately. We observe that AUC of CVs remains below the baseline whatever the percentage of compression. We notice however a gradual decrease in AUC scores depending on the percentage of compression. We explain this by the fact that a CV is already a summary of the most important information about the candidates and thereby attempting to summarize degrades final results. We apply the same process with cover letters. Performance is still low overall for CLs in comparison with CVs, however, there is a slight increase in AUC scores with a compression rate of 30%. We explain these results by particular information contained in a cover letter such as the form of address at the beginning or end of the letter (see Fig. 4) which are noise for the ranking system of E-Gen. Results with TMP segmentation (i.e. conserving only Title, Mission, and Profile of job offer) are of better quality.

Table 3 presents the results obtained by combining both parts of the application. Full text values are computed with the whole documents of the application. The first two columns show the results obtained by combining the summary of the CV and the CL. We observe again a deterioration in the results when trying to summarize the CV. Even if results are lower, it should be noted, however, that the best score is again obtained at 30%. The last two columns present the results with a summarized CL and the full CV. We observe an overall improvement of the AUC score and the best results with a compression rate of 30% of the Cover Letter.

Next step is to combine summaries of the cover letter, which suppresses noise and enriches the offer with the Relevance Feedback process. Table 4 presents the results obtained with different sizes of Relevance Feedback (RF1 corresponds to one application added to the job offer, RF2 two applications added to the job offer, etc.). Each application added with the relevance feedback process consists in a full CV and a summary of the cover letter with a compression rate of 30%. A random distribution of applications produces an AUC approximately at 0.5 like explained in Fawcett (2006). We compare *ISMIS Result* with those obtained using a summary of the cover letter. Each test is carried out 100 times with a random distribution of **relevant** applications for Relevance Feedback. Then we compute an average of AUC scores obtained (the curve shows the

Table 4
Comparison of AUC score for each size of Relevance Feedback with CORTEX summarization system.

Size of Relevance Feedback	*ISMIS result*	Full CV and CL summary 30% compression rate
Random distribution	0.500	0.500
RF0	0.642	0.653
RF1	0.654	0.658
RF2	0.657	0.659
RF3	0.659	0.661
RF4	0.659	0.659
RF5	0.660	0.662
RF6	0.661	**0.663**

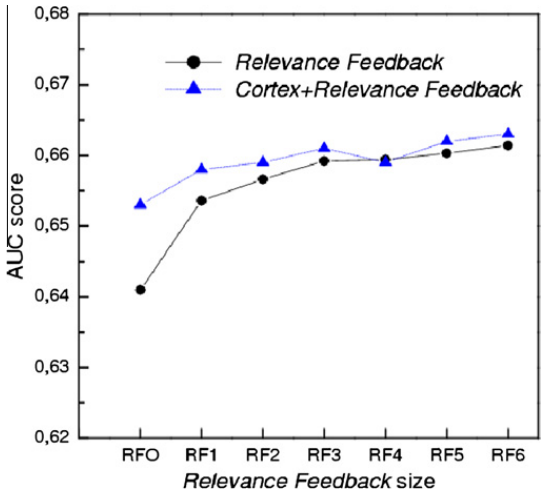

Fig. 5. Results of Relevance Feedback with and without summaries of CL.

average for each size). In fact, we compute the Residual Ranking (Billerbeck & Zobel, 2006): Documents that are used for Relevance Feedback are removed from the collection before ranking with the reformulated query. We assume that the Relevance Feedback process would behave as a reinforcement learning (Sutton & Barto, 1998) but it is impossible to experiment RF*n* with $n > 6$ with this corpus because the number of **relevant** candidates is too small for some job offers (see Table 1). We observe a slight improvement in results for almost any size of Relevance Feedback. We are conscious that the performance gain is low, however, it confirms previous results on the Cover Letter. Fig. 5 shows this improvement. This figure confirms that the addition of just one relevant candidate (RF1) enables the AUC value to be enhanced (i.e. an improvement of 0.5–1.2%). This Relevance Feedback (i.e. RF1) is not very time-consuming for the expert.

Fig. 6 shows detailed results of one test. For clarity reasons, we present only 3 of the 12 jobs of our dataset in order to compare results with and without CORTEX (for each job, *RFC* are AUC scores with CORTEX and *RF* without CORTEX).

For standard system, we observe a positive progress from 1% to 10% for 10 jobs between RF0 and RF1 (e.g. five jobs have an improvement between 5% and 10%). Note that between RF0 and RF6, 6 jobs have a significant positive progress between 10% and 12%. The combination of the E-Gen and CORTEX systems improve standard system results for five jobs from 1% to 5% between RF0 and RF1. Between RF0 and RF6, the Cortex version improves E-Gen's results for eight jobs from 1% to 5%.

The study of the results shows that job offer 31702 contains some relevant applications with a bad labeling (CV are labeled CL and CL are only a hyperlink to a CV). The reduction of information on the main document of the application leads the system version using summaries to degrade the AUC scores. Job offer 34861 shows a good improvement with each size of relevance feedback (RF0:0.65, RF1:0.70, RF6:0.73) and with CORTEX (RF0:0.68, RF1:0.72, RF6:0.79). The detailed study of results shows that job offer 33746 contains some empty applications labeled relevant. This leads the system with and without CORTEX to degrade final results. In the same way, an application added without CL explains the identical score in RF2 between RF and RFC for job offer 31274.

6. Conclusion and future work

Job offer processing is a difficult and highly subjective task. The retrieval of relevant information concerning job descriptions and skills is not a trivial task (Loth et al., 2010) and results on this type of document have been quite low (Clech &

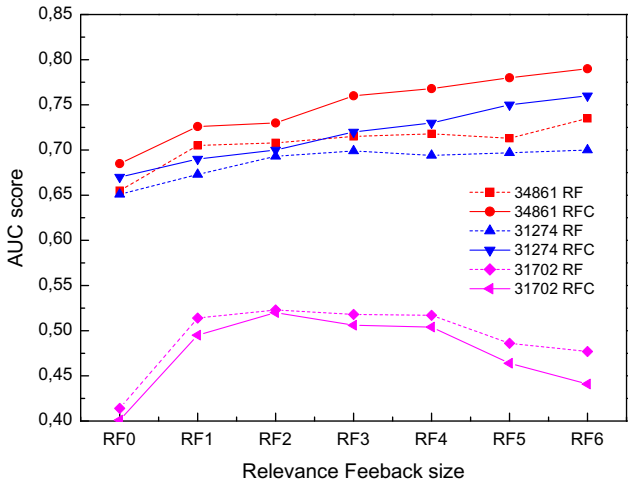

Fig. 6. Comparison of detailed results for 3 jobs with and without summaries of CL. For each job, *RFC* means AUC scores with CORTEX and *RF* without CORTEX.

Zighed, 2003). The information we use in this kind of process is not well formated in natural language, but follows a conventional structure. This paper deals with the CORTEX summarizer and the E-Gen system for processing job offers. E-Gen assists an employer in the recruitment task. This paper focuses on candidate answers to job offers. We rank the candidate answers by using different similarity measures and different document representations in a vector space model. We use a process of relevance feedback to perform reinforcement learning, whereby each new application added to the process assists in the decision-making. We choose to evaluate the quality of our approaches by computing *Area Under the Curve*. CORTEX is a summarization system using an optimal decision algorithm that combines several metrics. We present the results obtained by combining both systems. AUC obtained with summarized cover letter at 30% of compression size and a full CV shows a slight improvement in the results. As future work, we plan to apply other techniques, such as finding discriminant features of irrelevant applications using the Rocchio algorithm (Rocchio, 1971), weighting the different parts of an application, etc. in order to improve results. We also plan to use a categorization of jobs to take into consideration similar jobs, such as "developer" and "programmer". Finally we propose to measure the CV quality by building an evaluation on an Internet portal. Our aim with this evaluation is to present a job-seeker with a list of the most suitable job ads according to his profile.

Acknowledgements

Authors thank Richard James, Véronique Moriceau, André Bittar, ANRT *(Agence Nationale de la Recherche Technologique)* and Aktor Interactive that partially supported this work.

References

Audras, I., & Ganascia, J.-G. (2006). Apprentissage du frantais langue TtrangFre et TALN: Analyses de corpus Tcrits a l'aide d'outils d'extraction automatique du langage. In J.-M. Viprey (Ed.), *8Fmes JournTes d'Analyse de DonnTes Textuelles* (pp. 67–78). Univ. de Franche ComtT, Besanton 2006.

Bellot, P., & El-Bèze, M. (2001). Classification et segmentation de textes par arbres de dTcision. In *TSI* (Vol. 20, pp. 107–134). HermFs.

Ben Abdessalem Karaa, W. (2009). Web-based recruiting: A framework for cvs handling. In *Second international conference on web and information technologies "ICWIT'09", kerkennah Island, Sfax, Tunisia, June 12–14* (pp. 395–406).

Bernstein, A., Kaufmann, E., Kiefer, C., & Bnrki, C. (2005). Simpack: A generic java library for similarity measures in ontologies. Tech. rep., University of Zurich Department of Informatics.

Billerbeck, B., & Zobel, J. (2006). Efficient query expansion with auxiliary data structures. *Information Systems, 31*(7), 573–584.

Boudin, F., & Torres Moreno, J. M. (2007). Neo-cortex: A performant user-oriented multi-document summarization system. In *CICLing* (pp. 551–562).

Bourse, M., LeclFre, M., Morin, E., & Trichet, F. (2004). Human resource management and semantic web technologies. In *ICTTA 2004 Damascus Syria* (pp. 641–642).

Cazalens, S., & Lamarre, P. (2001). An organization of internet agents based on a hierarchy of information domains. In *Proceedings MAAMAW'2001, Annecy, France* (pp. 573–584).

Clech, J., & Zighed, D. A. (2003). Data mining et analyse des cv: une expérience et des perspectives. In *EGC'03 Revue des Sciences et Technologies de l'Information* (Vol. 17, pp. 83–92). Lyon.

Colucci, S., Di Noia, T., Di Sciascio, E., Donini, F. M., Mongiello, M., & Mottola, M. (2003). A formal approach to ontology-based semantic match of skills descriptions. *Journal of Universal Computer Science, Special issue on Skills Management, 9*, 1437–1454.

Dorn, J., & Naz, T. (2007). Meta-search in human resource management. In *Proceedings of 4th international conference on knowledge systems ICKS'07 Bangkok,Thailand* (pp. 105–110).

Enrica, A., & Iezzi, D. F. (2006). Recruitment via web and information technology: A model for ranking the competences in job market. In *JADT'2006, Besanton, France* (pp. 79–88).

Fan, R.-E., Chen, P.-H., & Lin, C.-J. (2005). Working set selection using the second order information for training SVM. *Journal of Machine Learning Research*, 1889–1918.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874.

Ferri, C., Flach, P., & Hernandez-Orallo, J. (2002). Learning decision trees using the area under the ROC curve. In *Proceedings of ICML 2002: Sydney, NSW, Australia* (pp. 139–146).

Gorenak, I., & Mlaker KaF, S. S. O. (2010). Cross-cultural comparison of online job advertisements. *JLST, Journal of Logistics and Sustainable Transport, 2*, 37–52.

Joachims, T. (1997). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML 1997, Nashville, Tennessee, USA* (pp. 143–151). San Francisco, CA, USA.
Kessler, R., Béchet, N., Roche, M., El-Bèze, M., & Torres-Moreno, J. M. (2008a). Automatic profiling system for ranking candidates answers in human resources. In *OTM '08 in Monterrey, Mexico* (pp. 625–634).
Kessler, R., Béchet, N., Roche, M., El-Bèze, M., & Torres-Moreno, J. M. (2009). *Job offer management: How improve the ranking of candidates*. Prague: ISMIS. 431–441.
Kessler, R., Torres-Moreno, J. M., & El-Bèze, M. (2007). E-Gen: Automatic job offer processing system for human ressources. In *MICAI, Aguscalientes, Mexique* (pp. 985–995).
Kessler, R., Torres-Moreno, J. M., & El-Bèze, M. (2008b). E-Gen: Profilage automatique de candidatures. In *TALN 2008, Avignon, France* (pp. 370–379).
Loth, R., Battistelli, D., Chaumartin, F., De Mazancourt, H., Minel, J. L., & Vinckx, A. (2010). Linguistic information extraction for job ads (SIRE project). In *RIAO'2010 9th conference 28–30 April, Paris, France* (pp. 300–303).
Marchal, E., Mellet, K., & Rieucau, G. (2007). Job board toolkits: Internet matchmaking and changes in job advertisements. *Human Relations, 60*(7), 1091–1113.
Mocho, M., Paslaru, E., & Simperl, B. (2006). Practical guidelines for building semantic e-recruitment applications. In *I-Know'06 special track on advanced semantic technologies, Graz, Austria, September 2006*.
Morin, E., LeclFre, M., & Trichet, F. (2004). The semantic web in e-recruitment. In *The first European symposium of semantic Web (ESWS'2004)* (pp. 67–78).
Quilan, J. (1993). *C4.5: Programs for machine learning*. San Mateo, CA, San Francisco, CA, USA: Morgan Kaufmann.
Rafter, R., Bradley, K., & Smyt, B. (2000). Automated collaborative filtering applications for online recruitment services. In *International conference on adaptive hypermedia and adaptive web-based systems, Trento, Italy* (pp. 363–368).
Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1994). Okapi at trec-3. NIST Special Publication 500-225: TREC-3, pp. 109–126.
Rocchio, J. (1971). Relevance feedback in information retrieval. In *The smart system: Experiments in automatic document processing* (pp. 313–323). Prentice-Hall.
Roche, M., & Kodratoff, Y., 2006. Pruning terminology extracted from a specialized corpus for CV ontology acquisition. In *OTM'06, Montpellier, France* (pp. 1107–1116).
Roche, M., & Prince, V. (2008). Evaluation et dTtermination de la pertinence pour des syntagmes candidats a la collocation. In *JADT* (pp. 1009–1020).
Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 288–297.
Salton, G., & Mcgill, M. J. (1986). *Introduction to modern information retrieval*. New York, NY, USA: McGraw-Hill Inc.
Smyth, B., & Bradley, K. (2003). Personalized information ordering: A case-study in online recruitment. *Journal of Knowledge-Based Systems*, 269–275.
Sparck Jones, K. (1970). Some thoughts on classification for retrieval. *Journal of Documentation*, 89–101.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction (adaptive computation and machine learning)*. The MIT Press.
Tolksdorf, R., Mocho, M., Heese, R., Oldakowski, R., & Christian, B. (2006). Semantic-Web-Technologien im Arbeitsvermittlungsprozess. *Wirtschaftsinformatik*, 17–26.
Torres-Moreno, J. M., Velázquez-Morales, P., & Meunier, M. (2001). CORTEX, un algorithme pour la condensation automatique de textes. In *ARCo* (Vol. 2, pp. 365–371).
Torres-Moreno, J. M., St-Onge, P.-L., Gagnon, M., El-Bèze, M., & Bellot, P. (2009). Automatic summarization system coupled with a question-answering system (qaas). In *CoRR* abs/0905.2990.
Torres-Moreno, J. M., Velazquez-Morales, P., & Meunier, J. (2002). Condensés de textes par des méthodes numériques. *JADT, St Malo, France, 2*, 723–734.
Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory, 13*, 260–269.
Yahiaoui, L., Boufaïda, Z., & Prié, Y. (2006). Semantic annotation of documents applied to e-recruitment. In *SWAP 2006 – Semantic web applications and perspectives*. ISSN: 1613-0073.