RESUMES : peRsonal knowlEdge baSe constrUction froM hEterogeneous Sources

Entreprise : Carian Software Developpment
Encadrement académique : Télécom SudParis
Laboratoire : SAMOVAR
Directrice de thèse : **Amel Bouzeghoub** amel.bouzeghoub@telecom-sudparis.eu
Co-encadrant de thèse : **Julien Romero** julien.romero@telecom-sudparis.eu
Encadrant au sein de l'entreprise : **Jean-Marie Volle**

1 Summary in English

The Web is composed of many documents of different nature, such as texts, images, or videos. These documents contain information about a wide range of topics that are noisy, unstructured, and ambiguous. Therefore, exploiting this variety is a huge challenge. When it comes to information about humans, one could use specialized websites such as social media, forums, blogs, or personal websites. However, it raises many problems. For example : How can we, from a single source, extract knowledge about a person ? How can we know that two accounts on two different websites represent a single person ? How does a person communicate with others ?

This kind of information can be valuable in many applications, and in particular for CV enrichment. Given a candidate's resume, we would like to complement it with external sources such as Linkedin, Reddit, or GitHub. These additional clues can help a recruiter to make the appropriate decisions.

This thesis aims to construct a Personal Knowledge Base (PKB) from information gathered online to complement a resume. A personal knowledge base is a collection of structured statements about a person that can be queried and on which one can reason.

For example, let's say we have a candidate called John. He has a GitHub page that we managed to link to his resume. We extracted statements such as "John, knows, Java" and "John, contributes to, Open Source projects" from his profile. These statements are now part of his PKB. Now, we find a StackOverflow account for the same username. This account answered many questions about Java. We might suppose that the two accounts belong to the same person, and therefore we can complete John's PKB. Suppose we know that this John is a potential candidate for a company working on open source projects written in

Java. In that case, we can boost his resume and present additional information to help the recruiter.

For this thesis, we will consider candidates with knowledge about several of the following skills :

— Fluent written and spoken English. Some knowledge of French can be useful.
— Machine/Deep Learning
— Natural Language Processing
— Programming and software development
— Information extraction
— Knowledge bases/Ontologies
— Logic and automated reasoning
— Semantic Web and Web crawling

This thesis is a CIFRE and a collaboration between Telecom SudParis and Carian Software Development. The position will start before October 2022.

2 Thesis Context And Scientific Challenges

The widespread use and prevalence of digital information sources of various types (image, video, text) have led to a rapid increase in data volume that contains valuable data, requiring advanced solutions to extract, organize, contextualize and handle heterogeneity for more effective use. These solutions are particularly beneficial in knowledge-driven decision support systems that need to gather multimodal information to provide actionable and explainable insights to decision-makers. In such systems, knowledge about a person - current situation, historical behavior, etc. - is central to enable personalized applications. This thesis project aims to automatically build a Personal Knowledge Base (PKB) for individuals from large-scale unstructured content.

The considered use case is recruitment management. Most recruiters select their candidates on the strength of their CV and choose the best candidates, thanks to a score computed by parsing the CVs. However, a resume is not enough to get a picture of the candidate's profile. In addition, much information in the resumes is not up to date and may contain incomplete or incorrect information, making it challenging to evaluate candidates.

This is where the scientific challenges of the thesis meet the application needs. Using a Personal Knowledge Base for data enrichment can significantly improve recruitment quality by creating meaningful relationships. Aggregating knowledge extracted from various sources (Linkedin, Reddit, GitHub, Stack Overflow, Kaggle, or YouTube) may complement candidates' resumes and build in real-time a synthetic view of professional and extra-professional characteristics. PKB will help the recruiters to qualify the candidate profile and make a choice quickly.

2.1 Scientific Challenges

As a result, we have identified the following four scientific key challenges to be addressed through this thesis project :

- *Knowledge extraction from multiple unstructured sources* : Extracting and storing information from unstructured sources remains a broad challenge that requires unsupervised and semi-supervised methods ;
- *Personal knowledge bases* : Building and populating Personal Knowledge bases is a complex task and relatively new compared to traditional knowledge bases construction.
- *Social network alignment* : Generally, users tend to behave differently in different social networks. It is challenging to discover a consistent hidden pattern from the behaviors. Social network alignment could rely on Personal Knowledge Graphs ;
- *Managing changing knowledge and behavior evolution* : User behavior evolves and is subject to uncertainty. Managing highly dynamic knowledge in graphs is not straightforward. A significant understanding of temporal constructs and change over time is needed to capture these variations. The model should be updated incrementally and efficiently as new data arrive.

3 Related Work

Recently, various e-recruitment approaches have been developed that exploit text processing, semantic, machine learning techniques, and feature extraction algorithms [1]. These approaches succeed in screening out irrelevant resumes but obtain low precision ratios for the similarity measure between CVs and the corresponding job offers and still suffer from the lack of semantic knowledge. [2] proposed a comparative analysis between existing e-recruitment systems and classification. This study reveals that the surveyed approaches focused on two main goals : (i) finding a tough match between job posts and resumes or (ii) ranking applicants' resumes according to their relevance to a given job post. This observation leads us to the assumption that no existing solution has focused on enriching resumes to better understand the candidates' profile and social behavior. To fulfill our requirements, we provide a brief review of the state-of-the-art related to personality prediction with social behavior and personal knowledge base.

3.1 Personality Prediction With Social Behavior

People often use social networks to state their views on topics that matter to them in different areas of interest. We can use these comments and interactions (like message content and type, interactions with friends and followers, influence, popularity, response time, etc.) to characterize the behavior and personality of the individual. Indeed, several studies [3, 4] show that there is a strong correlation between users' personality and their behavior on social media.

Personality evaluation in various research studies can be defined by five dimensions known as the "Big Five" (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience). They are primarily based on the Trait theory that studies personality in the field of Psychology. In the literature, many approaches are used for personality prediction : questionnaires, Internet and online social networking site usage, linguistic techniques, or social network analysis based on links. However, all these approaches have revealed their limits to deduce someone's personality.

On the other hand, user behavior analysis extracts hidden user activity to provide additional information about users' profiles. By integrating this social activity information from multiple sources, more comprehensive knowledge about users can be achieved. As stated in [5], social network alignment and behavior analysis can benefit from each other. The inferred links can help information propagation across various social networks. Hence, social network alignment can help the behavior analysis of a person among multiple online social networks.

3.2 Personal Knowledge Bases

A personal knowledge base or graph (PKB) is a "source of structured knowledge about entities and the relation between them, where the entities and the relations between them are of personal, rather than general, importance." [6]. They are centered around a special node connected (directly or not) to all other nodes : the user. Apart from this fact, the knowledge representation is similar to traditional knowledge bases, with potential time-dependent statements. In the case of PKBs, the information we gather may be sparse and therefore harder to exploit for some algorithms. For example, entity linking (linking an entity in a text with an entity in a knowledge base) becomes challenging as we have to deal with a long tail of sparse entities.

The question of populating these PKBs is also open. There are three ways to do it. The most straightforward approach is to fill a PKB manually by asking targeted questions to the user. Some systems like Monica allow the users to record personal information about the people they know. In this case, we also talk about personal CRM (Customer Relationship Management). Some approaches we present below build a small PKB for training purposes by asking the participating users to fill in personal information before providing a sample of their conversations or tweets.

The second approach consists of using traditional information extraction techniques to extract what we require. Although that might work with clean sources like Wikipedia, it becomes harder for noisier sources that we will get to construct PKBs. Indeed, we are more likely to retrieve information from social media or personal blogs, far less structured than an encyclopedia. In [7], the authors use Freebase [8] and the search engine Bing to retrieve web snippets containing personal information (like the date of birth or the children) of people in Freebase. Then, they train a model to predict personal details. However, that supposes that the person is important enough and that the web snippet actually contains the information.

The third approach infers users' attributes directly from a text, without the information being explicitly mentioned in the text. So, for example, we could try to guess the age, sex, or job of a person directly from their tweets.

Some information can be inferred directly from our online traces. For example, [9] shows that personal traits and attributes like age, use of alcohol, or marital status can be deduced from Facebook likes or simple message [10]. [11] used a deep neural network to infer the age, gender, profession, and family status of a person from conversations (in movies scripts or online). For online content, they automatically generated a dataset using high precision patterns to extract a ground truth. This was possible thanks to the high amount of data available.

The personality of a person might be of interest in a PKB. However, it can be hard to infer, even for humans. [12] shows that computers are better than humans to guess personality traits. They used a simple linear regression and predicted OCEAN scores from Facebook likes.

Interpersonal information is also essential to describe a person. For example, in the case of recruitment, we would like to guess how someone will integrate with a team. PRIDE [13] infers relationships between participants in a conversation (like friend, boss, mother) using a neural network that can deal with multiple speakers. [14] focuses on characterizing the nature of an interaction, for example, to say if a speaker is cooperative, active, or distant. They mainly provided a dataset for this task and tested a simple SVM model.

Although much work has been proposed, we believe there is still room for improvement to achieve our goal.

4 Impact of the Results

We will promote the thesis results across various application domains (including intelligent personal assistants and recommender systems). In particular, the results are of utter importance to Carian Software Development (CSD) for recruitment management, as recruiting is costly, but a data-driven approach can bring tangible improvements. However, hiring job applicants based on obsolete, inaccurate, or incomplete data prevents confident decision-making recruitment.

CSD promotes itself as a framework to enable smart hiring, and this goes by providing actionable datasets. Moving from manual resumes analysis to smart recruiting, relying on AI and big data to offer recruiters and decision-makers actionable information that helps them recruiting talents more effectively, regardless of the environment, opens the door to breakthrough innovation. In that context, CSD expects a rapid take-off if the project succeeds. The goal for CSD is to integrate the results of this thesis to the core of its data platform, as it will provide a strong differentiator and value to any company willing to develop innovative services.

5 Expected Contributions

- T1 : Definition/choice of external sources and a benchmark and handling of CSD platform ;
- T2 : Consolidation of the state-of-the-art in the domain of Personal Knowledge Graphs, e-recruitment, Social Network Analysis ;
- T3 : Proposition of a model for personal knowledge representation ;
- T4 : Proposition of a solution for knowledge extraction and personal knowledge graphs population ;
- T5 : Development of a unified framework for end-to-end resumes enrichment for decision making ;
- T6 : Evaluation of the proposals on the selected dataset.

Références

[1] J. Rout, S. Bagade, Y. Pooja, and N. Patil, “Personality evaluation and cv analysis using machine learning algorithm,” *International Journal of Computer Sciences and Engineering*, vol. 7, pp. 1852–1857, 5 2019. [Online]. Available : https://www.ijcseonline.org/full_paper_view.php?paper_id=4494

[2] M. Maree, A. B. Kmail, and M. Belkhatir, “Analysis and shortcomings of e-recruitment systems : Towards a semantics-based approach addressing knowledge incompleteness and limited domain coverage,” *Journal of Information Science*, vol. 45, no. 6, pp. 713–735, 2019.

[3] S. Dhelim, N. Aung, M. A. Bouras, H. Ning, and E. Cambria, “A survey on personality-aware recommendation systems,” *CoRR*, vol. abs/2101.12153, 2021. [Online]. Available : https://arxiv.org/abs/2101.12153

[4] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, “Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging,” *J. Big Data*, vol. 8, no. 1, p. 68, 2021. [Online]. Available : https://doi.org/10.1186/s40537-021-00459-1

[5] F. Ren, Z. Zhang, J. Zhang, S. Su, L. Sun, G. Zhu, and C. Guo, “Banana : when behavior analysis meets social network alignment.” in *IJCAI*, 2020, pp. 1438–1444.

[6] K. Balog and T. Kenter, “Personal knowledge graphs : A research agenda,” in *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, 2019, pp. 217–220.

[7] X. Li, G. Tur, D. Hakkani-Tür, and Q. Li, “Personal knowledge graph population from user utterances in conversational understanding,” in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 224–229.

[8] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase : a collaboratively created graph database for structuring human knowledge,”

in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.

[9] M. Kosinski, D. Stillwell, and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior,” *Proceedings of the national academy of sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.

[10] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, “Personality, gender, and age in the language of social media : The open-vocabulary approach,” *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[11] A. Tigunova, A. Yates, P. Mirza, and G. Weikum, “Listening between the lines : Learning personal attributes from conversations,” in *The World Wide Web Conference*, 2019, pp. 1818–1828.

[12] W. Youyou, M. Kosinski, and D. Stillwell, “Computer-based personality judgments are more accurate than those made by humans,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 4, pp. 1036–1040, 2015.

[13] A. Y. Anna Tigunova, Paramita Mirza and G. Weikum, “Pride : Predicting relationships in conversations,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[14] F. Rashid and E. Blanco, “Characterizing interactions and relationships between people,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels, Belgium : Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4395–4404. [Online]. Available : https://aclanthology.org/D18-1470