English version on page 3.

Représentation de données prosopographiques par des graphes de connaissance

Contexte scientifique

Le projet ANR DAPHNE (**D**écouverte dans les b**A**ses **P**rosopographiques **H**istoriques de co**N**-naissanc**E**s)[1] vise à formaliser les connaissances sur lesquelles s'appuient les historiens dans le cadre de données prosopographiques (reconstitution de carrières de personnages historiques), ainsi qu'à caractériser la dimension calculatoire du processus d'élaboration et de validation des hypothèses. Le projet associe étroitement des partenaires académiques spécialistes de l'extraction de motifs, de l'interrogation de bases de données, et de la qualité des données dans les systèmes d'information, ainsi que des historiens faisant de l'analyse prosopographique des universitaires des XIII^{e} au XV^{e} siècles.

Le projet Studium Parisiense[2] contient plus de 17 000 fiches recensant des carrières d'universitaires. Celles-ci sont constituées de *factoïdes*, des proto-faits attestés par des sources. Quand un factoïde atteint un certain degré de fiabilité (par exemple, parce que plusieurs sources le corroborent), on parle de *fait*. Ces données ont été saisies sur une longue période, avec des règles de formatage correspondant au travail historique.

Les premières étapes du projet se sont concentrées sur l'élaboration d'un modèle de données capturant les concepts principaux du travail de prosopographie historique, et un ensemble de règles pour qualifier la fiabilité des sources et la crédibilité des informations [1].

Description du travail

Dans un premier temps, le post-doctorat consistera à finaliser la modélisation, en vue de proposer une méthode de représentation et de gestion des données adaptée à la production et à la validation de connaissances historiques. Nous envisageons pour cela le développement d'une ontologie et l'utilisation de graphes de connaissance incertains (*Uncertain Knowledge Graphs* ou UKG) [2, 3]. Des formalismes adaptés, comme RDF, pourront compléter le travail. Les données pourront être enrichies par une liaison à des bases sémantiques (Yago) ou par une transformation permettant leur intégration dans un processus NLP[3] classique (extraction d'entités nommées).

L'objectif est de permettre une manipulation avancée des éléments composants les carrières des universitaires, c'est-à-dire avec des opérations de filtrage, de sélection, de restriction ainsi que des alignements ou changements d'échelle sur l'axe temporel, ou de la composition de requêtes. Dans un second temps, nous espérons pouvoir proposer des rapprochements nouveaux grâce à de l'inférence. La prise en compte de la qualité des données est aussi importante, afin d'une part de considérer l'imprécision des données (aussi bien dans le temps que l'espace), et d'autre part d'intégrer au calcul

1. Voir `http://daphne.huma-num.fr/` et `https://anr.fr/Projet-ANR-17-CE38-0013`
2. `http://studium.univ-paris1.fr/`
3. *Natural Language Processing*

un mécanisme d'estimation de la qualité du résultat, tenant compte des indicateurs de qualité des données initiales et des opérateurs spécifiques utilisées et combinés dans la requête.

L'équipe Vertigo[4], au sein de laquelle seront réalisés les travaux du post-doctorat, a proposé des travaux similaires dans le contexte de la notation musicale (synthétisés dans [4]). L'intégration de l'incertitude et de la qualité représentent un défi supplémentaire.

Le deuxième aspect du travail repose en partie sur le premier, puisqu'il s'agit de proposer des méthodes de fouilles de données (*data mining*) adaptées aux données, par exemple par de la recherche de motifs dans les *Uncertain Knowledge Graphs* obtenus. Le temps disponible pour ce deuxième axe dépendra de la qualité des modélisations et du rythme de travail.

Profil

- doctorat/PhD en Informatique
- bon niveau en bases de connaissances (ontologies, *knowledge graphs*, XML)
- bonnes compétences en programmation (Java, Python)

Ne vous auto-censurez pas si vous n'avez pas précisément toutes les compétences énoncées.

Lieu de travail, rémunération, dates

- Les travaux de recherche se dérouleront dans l'équipe Vertigo du laboratoire CÉDRIC, sous la direction de Raphaël Fournier-S'niehotta, Nicolas Travers et Philippe Rigaux. Les locaux sont situés dans le 3e arrondissement de Paris. En raison des mesures sanitaires en vigueur fin 2020, il est probable qu'une partie au moins du post-doctorat se déroulera à distance.
- La durée du contrat est de 12 mois, avec un début possible courant mars 2021.
- Une rémunération mensuelle de 2500 euros environ.

Contact

Merci d'envoyer votre demande accompagnée d'un CV aux adresses :
fournier@cnam.fr, philippe.rigaux@cnam.fr et nicolas.travers@devinci.fr.

Références

[1] J. Akoka, I. Comyn-Wattiau, S. Lamassé, and C. du Mouza. Contribution of conceptual modeling to enhancing historians' intuition - application to prosopography. In *Conceptual Modeling - 39th International Conference, ER*, 2020.

[2] M. W. Chekol, G. Pirrò, J. Schoenfisch, and H. Stuckenschmidt. Marrying uncertainty and time in knowledge graphs. In S. P. Singh and S. Markovitch, editors, *AAAI*, pages 88–94, 2017.

[3] X. Chen, M. Chen, W. Shi, Y. Sun, and C. Zaniolo. Embedding uncertain knowledge graphs. In *AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*, pages 3363–3370, 2019.

[4] R. Fournier-S'niehotta, P. Rigaux, and N. Travers. Modeling Music as Synchronized Time Series : Application to Music Score Collections. *Information Systems*, 73:35–49, Mar. 2018.

4. https://cedric.cnam.fr/lab/equipes/vertigo/

Prosopographic data representation with knowledge graphs

Scientific context

The DAPHNE ANR project [5] aims at formalizing knowledge used by historians for prosopographic analyses (such as careers of historical figures), and at precising the conditions in which parts of the hypothesis development and validation may be automatized. The project tightly associates academic partners specialists in pattern mining, database querying, data quality in information systems, as well as historians studying academics from the 13th to 15th centuries.

The Studium Parisiense [6] project contains more than 17,000 records with careers of academic figures. Those records are based on *factoids*, *i.e.,* proto-facts mentioned by sources. When a factoid reaches a certain degree of reliability (for instance, because several sources support it), it is called a fact. Those data were recorded over a long period, with formating rules adapted to historical work.

The first steps of the project focused on elaborating a data model to capture the main concepts from prosopographic work, and a set of rules to rate the reliability of sources and the credibility of information [1].

Project description

The post-doctoral mission will first consist in finalizing the modelling, in order to propose a data representation and management method suitable to produce and validate historical knowledge. We envision developing an ontology and employing Uncertain Knowledge Graphs (UKGs) [2, 3]. Specific formalisms such as RDF could complete the toolbox. We may then enrich data by linking them to semantic databases (Yago) or by transforming them to integrate a classical NLP workflow (with Named Entity Recognition).

The goal is to enable advanced manipulation of careers and their constituting elements, *i.e.,* filtering, selecting, grouping, as well as aligning them on different timescales. Then, we want to use inference to propose new links between records. Data quality is also important, since considered data are imprecise (in both temporal and spatial dimensions) and because we are willing to integrate an estimation of the quality of the results into querying/management methods, based on initial data quality and the precision of the operators combined in the query.

The Vertigo team [7], in which the candidate will be recruited, has already proposed similar works in the context of musical notation (see a synthesis in [4]). However, integrating data quality and uncertainty is a new challenge.

The second aspect of the work relies on the first one : we aim at introducing data mining methods adapted to these data, for instance mining patterns in the obtained UKGs. The time allocated to this task will depend on the quality of the modeling work, and its pace.

5. **D**écouverte dans les b**A**ses **P**rosopographiques **H**istoriques de co**N** naissanc**E**s, see `http://daphne.huma-num.fr/`.
6. `http://studium.univ-paris1.fr/`
7. `https://cedric.cnam.fr/lab/equipes/vertigo/`

Candidate background

- PhD in computer science
- good level with knowledge databases (ontology, knowledge graphs, XML)
- proficient programming skills (Java, Python)

Feel free to apply even if you do not have precisely all the mentioned skills.

Workplace, dates, salary

- The candidate will be integrated in the Vertigo team (CÉDRIC lab at CNAM Paris), under the superivision of Raphaël Fournier-S'niehotta, Nicolas Travers and Philippe Rigaux. Offices are located in the center of Paris. Due to the ongoing pandemic, the beginning of the contract (at least) is likely to be remote.
- One-year contract, with a starting date in march 2021, if possible.
- The net salary will be around 2,500 euros per month.

Contact

Please send your application with a resume at the following addresses : fournier@cnam.fr, philippe.rigaux@cnam.fr et nicolas.travers@devinci.fr.

Références

[1] J. Akoka, I. Comyn-Wattiau, S. Lamassé, and C. du Mouza. Contribution of conceptual modeling to enhancing historians' intuition - application to prosopography. In *Conceptual Modeling - 39th International Conference, ER*, 2020.

[2] M. W. Chekol, G. Pirrò, J. Schoenfisch, and H. Stuckenschmidt. Marrying uncertainty and time in knowledge graphs. In S. P. Singh and S. Markovitch, editors, *AAAI*, pages 88–94, 2017.

[3] X. Chen, M. Chen, W. Shi, Y. Sun, and C. Zaniolo. Embedding uncertain knowledge graphs. In *AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*, pages 3363–3370, 2019.

[4] R. Fournier-S'niehotta, P. Rigaux, and N. Travers. Modeling Music as Synchronized Time Series : Application to Music Score Collections. *Information Systems*, 73:35–49, Mar. 2018.