Graphe de connaissances et folksonomie : leur performance comparative dans le calcul de l'affinité

Chun Lu[1, 2], Philippe Laublet[1], Milan Stankovic[1, 2] et Filip Radulovic[2]

[1] Laboratoire STIH, Université Paris-Sorbonne, 28 rue Serpente, 75006 Paris
philippe-laublet@paris-sorbonne.fr
[2] Sépage, 27 rue du Chemin Vert, 75011 Paris
{chun, milstan, filip}@sepage.fr

Résumé : L'affinité est un élément essentiel dans bien des systèmes d'information centrés sur l'utilisateur comme les systèmes de recommandation. Le graphe de connaissances et la folksonomie sont respectivement des jalons importants pour le Web Sémantique et le Web Social. Nonobstant leur trait collaboratif partagé (du moins quelques grands graphes de connaissances le sont), les données codées diffèrent tant par la nature (fait versus expérience) que par la structure (formelle versus lâche). Dans ce papier, nous tentons d'éclaircir leur performance comparative dans la tâche du calcul de l'affinité à travers deux expériences dans le domaine du e-tourisme. Nos résultats montrent que le graphe de connaissances permet de calculer l'affinité avec plus de précision alors que la folksonomie augmente la diversité et la nouveauté. Ces constatations nous ont motivés à développer le Framework d'Affinité Sémantique pour bénéficier de leurs avantages respectifs. L'original de ce papier est publié par ESWC 2017.

Mots-clés : Affinité, recommandation, collaboratif, similarité, sémantique, graphe de connaissances, folksonomie, e-tourisme

1 Introduction

L'affinité entre un utilisateur et une entité (ex. film, musique, artiste) est la probabilité que l'utilisateur soit attiré par l'entité ou réalise une action liée à elle (ex. cliquer, acheter, aimer, partager). L'affinité est un élément essentiel dans bien des systèmes d'information centrés sur l'utilisateur comme les systèmes de recommandation, la publicité en ligne, la recherche exploratoire etc. Parmi les techniques de calcul de l'affinité, celles basées sur le contenu posent l'hypothèse que l'utilisateur aurait une affinité plus élevée avec les entités similaires à celles qu'il a appréciées dans le passé. Le graphe de connaissances et la folksonomie sont respectivement des jalons importants pour le Web Sémantique et le Web Social. Ils ont tous deux boosté les techniques basées sur le contenu grâce au grand nombre de données disponibles sur les entités. Sur le Web Sémantique, les utilisateurs contribuent à la création des graphes de connaissances universels comme DBpedia et Wikidata. Sur le Web Social, les utilisateurs annotent et catégorisent les entités avec des étiquettes libres formant des folksonomies. Nonobstant leur trait collaboratif partagé, les données codées diffèrent tant par la nature que par la structure. Les graphes de connaissances structurent des données factuelles avec une ontologie. Les folksonomies contiennent des données d'expérience avec une structure lâche. Nous donnons un exemple pour illustrer leur différence. Sur DBpedia, le film *dbr:Jumanji* est associé aux faits comme *dbr:Joe_Johnston* par la propriété *dbo:director*, *dbr:Robin_Williams* par *dbo:starring*. Dans la folksonomie de MovieLens[1], le même film est abondamment annoté avec des étiquettes comme « nostagic », « not funny », « natural disaster » etc. Ces étiquettes reflètent l'expérience qu'ont eue les différents utilisateurs et ainsi une sorte d'intersubjectivité qui n'est pas présente dans les graphes de connaissances.

[1] https://movielens.org/

Après une étude approfondie de la littérature (section 2), nous n'avons pas réussi à trouver des éclairages utiles sur l'efficacité comparative de ces deux espaces de données dans la tâche du calcul de l'affinité. Ces deux espaces de données continuant de proliférer sur le web, il est plus que jamais nécessaire de faire la lumière sur cette question. Nous étudions cette question à travers deux expériences dans le domaine du e-tourisme. La première expérience hors ligne est décrite dans la section 3 dont les constatations ont motivé le développement du Framework d'Affinité Sémantique (section 4). La section 5 présente la deuxième expérience qualitative et évalue le framework proposé. La section 6 conclut le papier.

2 Travaux connexes

Depuis plus d'une décennie, les chercheurs étudient de près les liens entre le Web Sémantique et le Web Social. L'idée générale derrière ces efforts est d'augmenter la sémantique du Web Social à l'aide des technologies du Web Sémantique (Bontcheva et Rout, 2014). (Passant et Laublet, 2008) proposent l'ontologie *MOAT* et un framework collaboratif pour guider les utilisateurs à fournir la sémantique des tags (étiquettes) lors du processus d'annotation. (Mika, 2007) propose de construire des ontologies légères à partir des folksonomies. (Cantador et al., 2011) présentent une méthode utilisant le Web Sémantique et le traitement automatique des langues pour classer les en fonction de l'intention derrière l'application des tags. Certains auteurs essaient d'extraire des préférences utilisateurs à partir des folksonomies (Orlandi et al., 2012). D'autres essaient de prouver l'avantage de les utiliser dans les systèmes de recommandation (Semeraro et al., 2012). Du côté des graphes de connaissances, les auteurs les exploitent pour calculer la similarité sémantique entre les entités et les incorporent dans les systèmes de recommandation. Dans (Passant, 2010) et (Piao et Breslin, 2016), les auteurs présentent respectivement *Linked Data Semantic Distance* et *Resource Similarity* qui exploitent DBpedia pour recommander des artistes musicaux. Des variantes de *Spreading Activation* sont utilisées dans des systèmes de recommandation inter-domaines (Kaminskas et al., 2014) et de recherche exploratoire (Marie, 2014).

3 Première expérience

Nous avons mené une première expérience dans un scénario de recommandation des destinations de voyage.

3.1 Jeu de données

Nous avons adapté le jeu de données Yahoo! Flickr Creative Commons 100[2] (YFCC100M) (Thomee et al., 2016). YFCC100M contient 100 millions de photos et de vidéos géo-localisées et datées publiées sur Flickr. Les traitements principaux sont les suivants : 1. Retenir seulement les photos/vidéos avec la meilleure précision de géolocalisation 2. Associer chaque photo/vidéo à un lieu d'intérêt dans un graphe de connaissances (Lu et al., 2016) 3. Trier par ordre chronologique. A l'issue de ces traitements, nous avons obtenu, pour chaque utilisateur, une séquence de voyages contenant les villes qu'il a visitées successivement. Le jeu de données final[3] utilisé dans cette expérience contient 3878 utilisateurs et 705 villes. Les séquences de voyages contiennent en moyenne 5,27 villes.

[2] http://webscope.sandbox.yahoo.com/

[3] Le jeu de données ainsi que certaines autres ressources mentionnées se trouvent à : https://bitbucket.org/sepage/semantic-affinity-framework

3.2 Traitements de la folksonomie et du graphe de connaissances

Nous utilisons une folksonomie collectée sur un site de voyage collaboratif. Elle contient 234 étiquettes sur 26,237 villes dans 154 pays. Nous l'avons modélisée dans un espace vectoriel de type *tag genome* (Vig et al., 2012) où les villes sont notées sur une échelle continue de 0 à 1 pour chaque étiquette. La mesure cosinus est utilisée pour calculer la similarité entre les villes.

Pour chacune des 705 villes dans le jeu de données, nous avons exécuté des requêtes SPARQL avec toutes les propriétés sélectionnées (TABLE 1). Les ressources liées par la propriété *skos:broader* sont assimilées à celles liées par *dct:subject*. Nous avons ensuite éliminé les ressources qui sont liées à une seule ville car elles ne contribuent guère au calcul de la similarité. 501,365 ressources sont initialement obtenues et 29,743 d'entre elles sont finalement retenues. Nous avons adopté la mesure de Jaccard dont l'efficacité a été démontrée dans une comparaison avec des mesures plus sophistiquées (*VsmSim, GbkSim, FuzzySim)* dans un scénario de recommandation musicale (Nguyen et al., 2015).

TABLE 1 – *Propriétés DBpedia sélectionnées pour calculer la similarité entre les villes*

Entrant		**Sortant**	
dbo:birthPlace	dbo:broadcastArea	dbo:isPartOf	dbo:part
dbo:location	dbo:nearestCity	dbo:country	dbo:twinTown
dbo:deathPlace	dbo:ground	dbo:timeZone	dbo:saint
dbo:city	dbo:foundationPlace	dbo:Mayor	dbo:district
dbo:capital	dbo:assembly	dbo:region	dct:subject
dbo:hometown	dbo:restingPlace	dbo:province	(skos:broader)
dbo:recordedIn	dbo:place	dbo:leaderName	
dbo:residence	dbo:locationCity		
dbo:headquarter			

3.3 Calcul de l'affinité

Etant donné un profil utilisateur contenant une liste de villes visitées dans le passé, le score d'affinité d'une ville candidate v_i est la moyenne des scores de similarité qu'elle a avec chacune des villes de son profil.

$$affinité(u, v_i) = \frac{\sum_{v_j \in profil(u)} sim(v_i, v_j)}{|profil(u)|} \qquad (1)$$

3.4 Protocole et métriques

Pour une séquence de voyages de *n* villes, les *n-1* premières villes constituent le profil utilisateur, la *n-ième* ville est considérée comme la vérité terrain. Chaque approche prend le profil utilisateur en entrée et génère trois listes de recommandations contenant respectivement 10, 20, 30 villes dans l'ordre décroissant des scores d'affinité. Dans le scénario de la recommandation, la capacité du calcul de l'affinité est reflétée par la précision. Nous utilisons deux métriques pour la mesurer : Succès (2) et Mean Reciprocal Rank (MRR) (3). Depuis quelques années, les chercheurs qui travaillent sur les systèmes de recommandation portent un intérêt particulier sur la diversité et la nouveauté. Ces deux qualités sont mesurées avec les formules (5) et (6). Par analogie avec (Di Noia et al., 2014), la diversité intra-liste (ILS) est calculée par rapport à deux propriétés : *dbo:country* et *dct:subject*. Nous utilisons les valeurs pagerank des ressources DBpedia comme les scores de popularité. Comme dans (Nguyen et al., 2015), nous considérons que 20% des villes ayant les meilleurs scores sont populaires.

$$Succès = \frac{\sum_{u \in U} rel_{g,u}}{|U|} \text{ où } rel_{g,u} = \begin{cases} 1, si\ vérité\ terrain\ g\ est\ dans\ top-N \\ 0,\ sinon \end{cases} \quad (2)$$

$$MRR = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank_u} \quad (3)$$

$$ILS_u@N = \sum_{i \in L_u^N} \sum_{j \in L_u^N} \frac{sim(i,j)}{|paires|} \quad (4) \qquad ILS@N = \frac{1}{|U|} \sum_{u \in U} ILS_u@N \quad (5)$$

$$Nouveauté@N = \frac{nombre\ de\ villes\ recommandées\ impopulaires}{N * |U|} \quad (6)$$

3.5 Résultats et discussions

Les résultats sont présentés dans la TABLE 2. Les tests t appariés montrent que les différences entre les deux approches sont statistiquement significatives sur toutes les métriques dans toutes les configurations. Nous observons un avantage net de GC sur FOLK en termes de succès et de MRR, ce qui reflète la capacité de détecter les villes qui sont en affinité élevée avec l'utilisateur et de mieux les ordonner. Les recommandations générées par FOLK sont plus diverses et plus nouvelles. Dans la folksonomie, certains aspects ne sont pas pas présents, tels que la géographie (*dbo:country, dbo:region)*, les personnes (*dbo:birthPlace, dbo:residence*), les catégories connexes (*dct:subject, skos:broader*). La folksonomie contient des traits tels que « Luxury Brand Shopping », « Clean Air » et « Traditional food ». Ces traits peuvent être partagés par de différentes villes du monde même si elles sont moins populaires. Ces constatations nous ont motivés pour développer le Framework d'Affinité Sémantique qui profite de la complémentarité des deux approches pour parvenir à un compromis équilibré sur la pertinence, la diversité et la nouveauté.

TABLE 2 – *Résultats de la première expérience, GC : Graphe de connaissance, FOLK : Folksonomie*

	Top-10		Top-20		Top-30	
	GC	FOLK	GC	FOLK	GC	FOLK
Succès	**0.232**	0.06	**0.33**	0.116	**0.386**	0.166
MRR	**0.047**	0.003	**0.047**	0.003	**0.047**	0.003
ILS	0.257	**0.089**	0.208	**0.072**	0.176	**0.065**
Nouveauté	0.717	**0.824**	0.722	**0.772**	0.723	**0.755**

4 Framework d'Affinité Sémantique

Nous proposons un Framework d'Affinité Sémantique (FAS) qui intègre, agrège, enrichit et nettoie les données sur les entités en provenance des graphes de connaissances et des folksonomies. Son pipeline est décrit dans FIGURE 1. Une explication plus détaillée se trouve dans (Lu et al., 2017). Un graphe d'affinité est généré à l'issue du processus. Nous avons également développé un mécanisme pour expliquer les recommandations. Par exemple, on peut expliquer la recommandation de *dbr:Ljubljana* par *dbc:Capitals_in_Europe*. Etant donné une liste d'entités, nous cherchons les caractéristiques les plus fréquentes tout en contrôlant leur diversité par le biais des propriétés qui relient les caractéristiques aux entités.

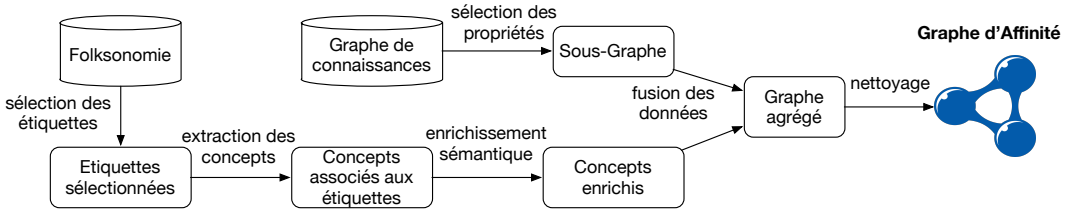

FIGURE 1 – Pipeline du Framework d'Affinité Sémantique

5 Deuxième expérience

Nous avons mené une deuxième expérience qualitative pour évaluer l'utilité et l'efficacité du Framework d'Affinité Sémantique. Nous avons généré un graphe d'affinité avec les données décrites dans 3.2 que nous nommons GA. A part la recommandation, nous nous sommes intéressés à la capacité d'explication des différentes approches.

You submitted:	You might like:	We recommend you:
dbr:Rome	dbc:Clothing	dbr:The_Hague
dbr:Florence	dbr:Food	dbr:Haarlem
dbr:Amsterdam	dbr:David_de_Haen	dbr:Naples
	dbr:Italy	dbr:Milan
	dbr:History	dbr:Turin

FIGURE 2 – Exemple de recommandations et d'explications générées par GA

37 personnes âgées de 25 à 38 ans ont participé à cette expérience dont 19 hommes et 18 femmes. Ils travaillent tous dans des sociétés sises dans une pépinière d'entreprise à Paris. Nous avons demandé aux participants de se mettre dans le scénario de rechercher pour la prochaine destination de voyage. Ils sont allés sur l'interface de test où les 705 villes étaient affichées dans un ordre aléatoire. Ils ont choisi plusieurs villes qui leur paraissaient intéressantes à première vue. Une fois que ces choix ont été soumis, ils ont reçu trois listes de recommandations contenant cinq villes accompagnées d'une explication avec cinq entités/étiquettes/catégories. FIGURE 2 montre un exemple généré par GA. Ils ont noté la pertinence, la diversité et la nouveauté/intérêt des recommandations et des explications sur une échelle de 1 à 5. Nous considérons que les notes supérieures à 3 comme notes positives. Nous utilisons comme métrique le pourcentage des notes positives.

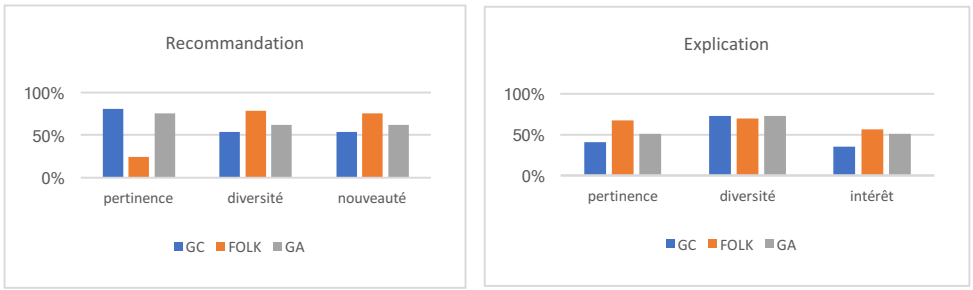

FIGURE 3 – Résultats de la deuxième expérience

Sur la recommandation, les résultats présentés dans FIGURE 3 sont en phase avec ceux de la première expérience. Sur l'explication, les résultats montrent que celle générée par FOLK est la mieux appréciée. Notre folksonomie a été créée de manière collaborative par les voyageurs. Elle couvre plusieurs aspects du voyage (nourriture, activité, transport). La capacité d'explication de GA a été boostée par l'inclusion de la folksonomie, ce qui l'a permis de surperformer GC. Les participants sont en général sceptiques envers GC. Certains trouvent ses explications assez générales, par exemple *dbr:Leisure*. D'autres les trouvent difficiles à comprendre, par exemple *dbr:China_Record_Corporation*. En effet, ces problèmes pourraient être résolues à travers l'utilisation de l'arbre des catégories DBpedia (définir un seuil en deça duquel une catégorie peut être considérée comme étant trop générale) ou le filtrage sur *rdf:type* (mettre certaines classes sur liste noire). Ces filtrages pourraient être intégrés dans le pipeline du Framework d'Affinité Sémantique.

6 Conclusion

Dans ce papier, nous avons étudié la performance comparative du graphe de connaissance et de la folksonomie dans la tâche du calcul de l'affinité à travers deux expériences dans le domaine du e-tourisme. Les résultats montrent que le graphe de connaissances permet de calculer l'affinité avec plus de précision alors que la folksonomie augmente la diversité et la nouveauté. Nous avons développé le Framework d'Affinité Sémantique pour bénéficier de leurs avantages respectifs. La combinaison des deux espaces de données aboutit à une performance équilibrée tant pour la recommandation que pour l'explication. Plus de détails se trouvent dans l'original de ce papier (Lu et al., 2017).

Références

Bontcheva, K., Rout, D. (2014). Making sense of social media streams through semantics: a survey. Semantic Web, 5(5), 373-403

Cantador, I., Konstas, I., Jose, J. M. (2011). Categorising social tags to improve folksonomy-based recommendations. Web Semantics: Science, Services and Agents on the World Wide Web, 9(1), 1-15

Di Noia, T., Cantador, I., & Ostuni, V. C. (2014, May). Linked open data-enabled recommender systems: ESWC 2014 challenge on book recommendation. In *Semantic Web Evaluation Challenge* (pp. 129-143). Springer International Publishing.

Kaminskas, M., Fernández-Tobías, I., Ricci, F., Cantador, I. (2014). Knowledge-based identification of music suited for places of interest. Information Technology & Tourism, 14(1), 73-95

Lu, C., Laublet, P., Stankovic, M. (2016). Travel attractions recommendation with knowledge graphs. In: Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management. Bologna, Italy

Lu, C., Stankovic, M., Radulovic, F., & Laublet, P. (2017, May). Crowdsourced Affinity: A Matter of Fact or Experience. In *European Semantic Web Conference* (pp. 554-570). Springer, Cham.

Marie, N. (2014). Linked data based exploratory search. Doctoral dissertation. Université de Nice Sophia-Antipolis

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. Web semantics: science, services and agents on the World Wide Web, 5(1), 5-15

Nguyen, P. T., Tomeo, P., Di Noia, T., & Di Sciascio, E. (2015). Content-based recommendations via DBpedia and Freebase: a case study in the music domain. In International Semantic Web Conference (pp. 605-621). Springer International Publishing.

Orlandi, F., Breslin, J., Passant, A. (2012). Aggregated, interoperable and multi-domain user profiles for the Social Web. In: Proceedings of the 8th International Conference on Semantic Systems (pp. 41-48). ACM

Piao, G., and Breslin, J. (2016). Measuring semantic distance for linked open data-enabled recommender systems. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing. ACM

Passant, A.: dbrec—music recommendations using DBpedia. In: Proceedings of the 9th International Semantic Web conference (pp. 209-224). Springer Berlin Heidelberg (2010)

Passant, A., Laublet, P. (2008). Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In: Proceedings of Linked Data on the Web workshop

Semeraro, G., Lops, P., De Gemmis, M., Musto, C., & Narducci, F. (2012). A folksonomy-based recommender system for personalized access to digital artworks. Journal on Computing and Cultural Heritage (JOCCH), 5(3), 11

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., and Li, L. J. (2016). YFCC100M: The new data in multimedia research. Communications of the ACM, 59(2), (pp. 64-73)

Vig, J., Sen, S., & Riedl, J. (2012). The tag genome: Encoding community knowledge to support novel interaction. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3), 13.