Estimer la notoriété d'un nom propre via Wikipedia

Mouna ELASHTER, Denis MAUREL
Université François-Rabelais de Tours, Laboratoire d'informatique
mouna.elashter@etu.univ-tours.fr, denis.maurel@univ-tours.fr

RÉSUMÉ

Cet article propose de calculer, via Wikipedia, un indice de notoriété pour les entrées du dictionnaire relationnel multilingue de noms propres Prolexbase. Cet indice de notoriété dépend de la langue et participera, d'une part, à la construction d'un module de Prolexbase pour la langue arabe et, d'autre part, à la révision de la notoriété actuellement présente pour les autres langues de la base. Pour calculer la notoriété, nous utilisons la méthode SAW (précédée du calcul de l'entropie de Shannon) à partir de cinq valeurs numériques déduites de Wikipedia.

ABSTRACT

Estimate the notoriety of a Proper name using Wikipedia.
This paper proposes to use Wikipedia for calculating the notoriety of the entries of Prolexbase, a multilingual relational dictionary of proper names. This notoriety is language dependent. It will be a first step in the construction of an Arabic module of Prolexbase, it also will take part to the notoriety revision currently present for the other languages in the database. To calculate the notoriety, we present a multi criteria technique, the method SAW (preceded by the calculation of Shannon entropy), starting from five numerical values deduced from Wikipedia.

MOTS-CLÉS : Notoriété, Noms propres, Prolexbase, Wikipedia.

KEYWORDS: Notoriety, Proper Name, Prolexbase, Wikipedia.

1 Introduction

La constitution d'un dictionnaire commence par le choix des entrées à y placer ou non. Cela est surtout vrai pour des dictionnaires édités sur papier, mais aussi pour des dictionnaires électroniques où certaines entrées peuvent augmenter inutilement l'ambiguïté. De plus, "il faut mentionner l'important problème que pose l'absence des noms propres dans les dictionnaires de langue" (Rey, 1977:30) et, pourrait-on ajouter aujourd'hui dans les dictionnaires électroniques. (Mikheev et al., 1999) constate que, dans la reconnaissance des entités nommées, "without gazetteers […] locations come out badly", mais il ajoute "The collection of gazetteers need not be a bottleneck: […] relatively small gazetteers are sufficient to give good Precision and Recall". En effet, les noms propres et leurs dérivés sont très souvent ambigus et une trop grande importance du dictionnaire peut créer des difficultés.

La recherche du mot *Paris* sur Geonames[1] donne des centaines de résultats, score à peine amélioré en précisant le type *city* (210 résultats). Sur Wikipedia[2], la même recherche conduit directement à la

[1] http://www.geonames.org/
[2] https://fr.wikipedia.org/

capitale de la France (qui est le plus connu des noms propres désignés par le mot *Paris*) et indique qu'il existe un grand nombre de pages homonymes. Paris (France) a plus de notoriété que d'autres villes du même nom. Bien sûr, d'autres critères spécifiques au corpus traité peuvent influencer les choix, mais un dictionnaire générique doit s'appuyer sur cette notion. L'évaluation de cette notoriété n'est cependant pas évidente. Doit-elle juste reposer sur le choix d'un *expert* ? Plusieurs travaux suggèrent l'utilisation du Web sémantique, et plus précisément de l'encyclopédie en ligne Wikipedia.

En particulier, le Macro Connections group au MIT Media Lab, qui travaille sur la quantification, l'analyse et la visualisation de la culture mondiale, a développé le projet Panthéon [Yu et al., 2016) qui calcule une notoriété universelle pour les personnes célèbres à travers l'histoire. Dans ce projet, on prend en compte le nombre de différentes versions linguistiques de Wikipédia ayant un article sur cette personne (en appliquant un filtre : ce nombre doit être supérieur à 25) et on calcule un *indice de popularité historique* (*historical popularity index*). Cette approche a été testée de plusieurs manières, entre autres par une comparaison avec le jeu de données de l'accomplissement humain (HA) ; HA est une compilation de 3 869 personnes notables dans les arts et les sciences, basée sur des encyclopédies imprimées : Panthéon contient 40 % des entrées disponibles dans HA, avec une corrélation significative entre les mesures d'impact historique dans Panthéon et dans HA.

Dans l'étude, sur vingt-quatre éditions linguistiques de Wikipedia, réalisée par (Eom, Shepelyansky, 2013) : à partir de la version anglaise, les auteurs ont extrait 1,1 million d'articles biographiques, puis ils ont cherché par les liens inter-langues les éventuelles pages correspondantes dans les vingt-trois autres éditions linguistiques et calculé un classement moyen basé entre autres sur le PageRank. En regardant les cinquante premiers de chaque édition, ils les répartissent en :

- des personnages historiques *mondiaux* apparaissant dans plus de dix-huit éditions linguistiques de Wikipedia (Linné, Platon, Napoléon…) ;
- et des personnages historiques *locaux* (Tycho Brahe, Sejong le Grand, Sun Yat-Sen…).

Citons aussi (Chevalier et al., 2010) qui s'intéresse non pas à la notoriété, mais à la qualité et à la fiabilité des articles de Wikipédia. Pour cela, les auteurs ont listé cinq indices :

- le nombre de mots ;
- le nombre de contributeurs et leur taux de contributions ;
- le nombre et taille des éditions ;
- le nombre de références et de liens internes ;
- la taille et l'activité de la discussion.

Enfin, présentons rapidement Prolexbase (Tran & Maurel, 2006). Il s'agit d'une ressource libre[3], un dictionnaire relationnel multilingue de noms propres, au format LMF (ISO 24613), depuis (Bouchou, Maurel, 2008). Au niveau multilingue se trouvent des pivots qui représentent un point de vue sur les noms propres et se projettent sur chaque langue en des prolexèmes, ensembles de formes morphosémantiquement liées (alias et dérivés). A chaque prolexème est associé, éventuellement, un lien vers Wikipedia et, obligatoirement, un indice de notoriété basé sur trois valeurs, conformément à la norme ISO 12620 : l'indice le plus fort étant 1 et le plus faible 3. Les indices actuels ont été choisis manuellement, ce que nous avons voulu changer, d'abord pour permettre l'ajout automatique de nouvelles entrées et de nouvelles langues, mais aussi pour autoriser une réévaluation régulière de cette notoriété, qui évolue dans le temps. Un premier travail d'ajout a été réalisé par (Savary et al., 2013) qui a fortement augmenté le nombre d'entrées polonaises en se basant sur le total des consultation de Wikipedia dans l'année précédant leur travaux. Actuellement, Prolexbase contient neuf langues, mais

[3] http://www.cnrtl.fr/lexiques/prolex/

principalement le français, l'anglais et le polonais qui seront seules considérées dans cet article. Le point de départ de ce travail était l'ajout de la langue arabe.

En préliminaire à notre recherche de la notoriété, qui sera basée sur les liens vers Wikipedia présents dans Prolexbase, nous avons dû consolider les URL que nous possédions. Comme nous avons travaillé sur trois langues, nous avons comparés les liens existants avec ceux obtenus à partir d'une redirection linguistique de Wikipedia. Si tous les liens obtenus pour un nom propre étaient identiques pour chacune des trois langues, ils étaient validés, sinon, confiés à un expert pour vérification manuelle. Si tout était correct, mais si un lien dans une langue manquait, il était ajouté sans supervision.

2 Le choix de critères

Comme dans les articles cités précédemment, nous avons commencé par choisir des critères, au nombre de cinq, puis nous avons appliqué un algorithme de calcul multicritères. Les trois premiers critères concernent l'article lui-même et les deux autres les liens de cet article avec les autres articles de Wikipedia et l'ensemble de la Toile. Précisons que nous avons fait un tel calcul en interne à une édition linguistique, considérant que la notoriété d'un nom propre dépend de la langue. Comme il nous semble peu intuitif de comparer la notoriété d'une célébrité à celle d'un lieu, nos calculs se font type par type. Pour cet article, nous nous focalisons sur le type *célébrité*.

Voici ces cinq critères :

1) le nombre de consultations de l'article ;
2) le nombre de contributeurs à l'article ;
3) la taille de l'article ;
4) le nombre de liens internes à Wikipedia pointant vers l'article ;
5) et le nombre de liens externes à Wikipedia contenus dans l'article.

Le premier indice ne se limite pas à une année, mais à l'ensemble des données disponibles sur le nombre de consultations mensuelles de l'article depuis 2008. Les deux indices suivants permettent d'estimer la fiabilité de l'article consulté et les deux derniers d'estimer son intégration dans Wikipedia et dans la Toile. Finalement, c'est la combinaison de ces cinq indices qui nous permettra de calculer la notoriété que nous attribuons au nom propre (section 4).

3 Le calcul des cinq indices

Comme dit ci-dessus, nous disposons pour le calcul du premier indice, via des services web, du nombre de consultations mensuelles de l'article depuis 2008. Un premier service[4] permet de connaître ce nombre à partir de décembre 2007 pour l'anglais et de février 2008 pour les autres langues et ce, jusqu'à fin 2015. Pour les consultations postérieures, nous utilisons un service web de Wikimedia[5] qui commence sa compilation en 2016.

[4] La concaténation http://stats.grok.se/fr/201501/Paris nous permet de connaître le nombre de consultations de l'article Paris dans l'édition française de Wikipedia en janvier 2015.

[5] De même, via https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article/fr.wikipedia/all-access/all-agents/Paris/daily/2016010100/2016013100, nous obtenons le nombre de consultations de l'article Paris dans l'édition française de Wikipedia en janvier 2016.

Nous avons choisi de prendre comme départ de notre compilation janvier 2008 et comme fin le dernier mois complet avant la date du jour. Dans cet article, nous considérerons la période 2008-2015 qui couvre 96 mois. Comme la notoriété d'un nom propre peut varier dans le temps, nous n'attribuons pas la même valeur à chaque mois, mais nous associons à un mois ce que nous avons appelé son *coefficient d'oubli*. Janvier 2008 sera associé au coefficient 1/96, février 2008 au coefficient 2/96, mars 2008 au coefficient 3/96, etc. Nous calculons donc pour chaque nom propre et pour chaque mois, le produit du nombre de consultations de son article par le coefficient d'oubli du mois en question. Finalement, nous faisons la somme de ces valeurs. C'est cette somme qui constituera notre premier indice.

Pour un article donné, Wikipedia fournit un service web[6] permettant d'obtenir la liste de tous ses contributeurs, et donc, leur nombre. Le même service permet d'obtenir la taille de l'article et le nombre de liens entrants. Le décompte des liens externes se fait directement à partir de l'article lui-même.

4 Le calcul de la notoriété

Nous avons donc obtenu pour chaque nom propre cinq indices de notoriété. Nous allons maintenant calculer une valeur finale, égale à 1, 2 ou 3. Pour cela, nous avons effectué un calcul multicritères, la méthode SAW (*simple additive weighting*) (Afshari, 2010) qui nécessite d'attribuer un poids à chaque critère. Ce poids est parfois défini arbitrairement par l'utilisateur. Nous avons préféré le déduire du calcul de l'entropie de Shannon (1948), comme, entre autres, (Safari et al., 2012) ou (Karami, Johansson, 2014).

4.1 Le calcul des poids de chaque critère

Pour commencer, pour chaque nom propre *i* et chaque critère *j*, nous normalisons les valeurs x_{ij} obtenues précédemment en une valeur c_{ij} comprise entre 0 et 1 ; si *m* est le nombre total de prolexèmes considérés dans une langue donnée :

$$c_{ij}=\frac{x_{ij}}{\sum_{i=1}^{i=m} x_{ij}} \text{ pour i} = 1..\text{m}, \text{j} = 1..5$$

Puis, nous calculons l'entropie E_j (comprise entre 0 et 1) :

$$E_j = \left(\frac{-1}{\ln(m)}\right)\sum_{i=1}^{i=m}\left[c_{ij}\ln(c_{ij})\right] \text{ pour j} = 1..5$$
$$\text{avec, par convention, } c_{ij}\ln(c_{ij}) = 0 \text{ pour } c_{ij} = 0$$

et le poids W_j de chaque critère :

$$W_j = \left(\frac{1-E_j}{\sum_{j=1}^{j=5}(1-E_j)}\right) \text{ pour j} = 1..5$$

[6] https://www.mediawiki.org/wiki/API:Main_page/fr

4.2 La méthode SAW

Nous commençons par multiplier chaque valeur normalisée c_{ij} par le poids W_j du critère correspondant et nous obtenons le score S_i d'un nom propre en sommant ces cinq valeurs :

$$S_i = \sum_{j=1}^{j=5} c_{ij} * W_j \text{ pour i} = 1..\text{m}$$

La répartition des prolexèmes entre les trois valeurs de notoriété n'est pas uniforme. Nous considérons qu'il y a un petit nombre de prolexèmes de notoriété 1, un nombre plus important de notoriété 2 et un grand nombre de notoriété 3 (la plus faible). Pour cela, nous attribuons tout d'abord la notoriété 1 aux prolexèmes de scores supérieurs à la moyenne plus l'écart-type de l'ensemble des scores :

$$\begin{cases} \bar{M} = Moyenne(S_i) \text{ pour i} = 1..\text{m} \\ \bar{E} = Ecart\_type(S_i) \text{ pour i} = 1..\text{m} \\ \quad Si\ S_i > \bar{M} + \bar{E}, N_i = 1 \end{cases}$$

Ensuite, nous attribuons la notoriété 2 aux prolexèmes de scores supérieurs à la moyenne plus la moitié de l'écart-type de l'ensemble des scores restants et la notoriété 3 aux autres prolexèmes.

$$\begin{cases} \bar{\bar{M}} = Moyenne(S_i) \text{ pour } \bar{S_i} \leq \bar{M} + \bar{E} \\ \bar{\bar{E}} = Ecart\_type(S_i) \text{ pour } \bar{S_i} \leq \bar{M} + \bar{E} \\ \quad Si\ \bar{M} + \bar{E} \geq S_i > \bar{\bar{M}} + \frac{1}{2}\bar{\bar{E}}, N_i = 2 \\ \quad\quad Si\ S_i \leq \bar{\bar{M}} + \frac{1}{2}\bar{\bar{E}}, N_i = 3 \end{cases}$$

5 Application

Nous avons testé notre calcul de notoriété comme annoncé. Le Tableau 1 donne les poids respectifs de chaque critère dans chaque langue et le Tableau 2 la répartition des prolexèmes suivant leur notoriété.

Tableau 1 : Les poids respectifs de chaque critère dans chaque langue

Critères	(1)	(2)	(3)	(4)	(5)
Français	0,157	0,055	0,276	0,296	0,216
Anglais	0,202	0,066	0,284	0,289	0,160
Polonais	0,111	0,097	0,290	0,219	0,284

Tableau 2 : La répartition des prolexèmes suivant leur notoriété

Notoriété	1	2	3	Total
Français	391	898	2536	3 825
Anglais	525	1170	2690	4 385
Polonais	385	1110	2905	4 400

Platon et Louis XIV sont parfaitement connus dans ces trois langues (Tableau 3) ; alors que Napoléon 1er est plus célèbre en français et en polonais qu'en anglais ; très connu en français, Antoine de Saint-

Exupéry l'est moins en anglais et encore moins en polonais ; quant à Aldo Moro, il a une légère notoriété française et une très faible en anglais et en polonais ; *a contrario*, Czeslaw Niemen et Czeslaw Kiszczak, respectivement très connu et connu en polonais, ne le sont guère en français et en anglais…

Tableau 3 : Comparaison des notoriétés entre des noms propres de trois langues

Nom propre	Notoriété		
	Français	Anglais	Polonais
Platon, Louis XIV	1	1	1
Napoléon 1er	1	2	1
Antoine de Saint-Exupéry	1	2	3
Aldo Moro	2	3	3
Czeslaw Niemen	3	3	1
Czeslaw Kiszczak	3	3	2

6 Conclusion

En utilisant principalement Wikipedia, nous avons pu associer automatiquement aux prolexèmes de Prolexbase un critère de notoriété. Comme attendu, la valeur de ce critère diffère suivant les langues. La prochaine étape pour l'ajout de l'arabe dans Prolexbase sera la création des prolexèmes correspondant à des liens Wikipedia entre les éditions françaises ou anglaises et l'édition arabe. Il s'agira d'extraire de l'article le nom propre correspondant, et éventuellement des alias ou dérivés, et de leurs associer leurs instances (ou formes fléchies).

Références

AFSHARI A., MOJAHED M., MOHD YUSUFF R. (2010). Simple Additive Weighting approach to Personnel Selection problem. *International Journal of Innovation, Management and Technology*. 1:5, 511-515.

BOUCHOU B., MAUREL D. (2008), Prolexbase et LMF: vers un standard pour les ressources lexicales sur les noms propres, *Traitement automatique des langues*, 49(1):61-88[7].

CHEVALIER, HUOT, FEKETE (2010). Visualisation de mesures agrégées pour l'estimation de la qualité des articles Wikipédia. *EGC 2010*.

EOM, SHEPELYANSKY (2013). Highlighting Entanglement of Cultures via Ranking of Multilingual Wikipedia Articles, *PLoS ONE*, 8(10):e74554.

ISO 12620:1999. *Computer applications in terminology - Data categories*.

ISO 24613:2008. *Language resource management - Lexical markup framework* (*LMF*).

[7] http://www.atala.org/-Varia,55-

KARAMI A., JOHANSSON R. (2014). Utilization of Multi Attribute Decision Making Techniques to Integrate Automatic and Manual Ranking of Options. Journal of Information Science and Engineering. 30:519-534.

MIKHEEV A., MOENS M., GROVER C. (1999), Named entity Recognition without Gazetteers, *EACL'99*:1-8.

REY A. (1977). *Le lexique : images et modèles. Du dictionnaire à la lexicologie*. Paris, Armand Colin.

SAFARI H., SADAT FAGHEYI M., SADAT AHANGARI S., REZA FATHI M. (2012). Applying PROMETHEE Method based on Entropy Weight for Supplier Selection. *Business Management & Strategy*. Vol. 3:1, 97-106.

SAVARY AGATA, MANICKI LESZEK, BARON MAŁGORZATA. 2013. Populating a Multilingual Ontology of Proper Names from Open Sources. Journal of Language Modelling. 1-2:189-225.

SHANNON C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, vol. 27. 379-423 et 623-656.

TRAN M., MAUREL D. (2006), Prolexbase : Un dictionnaire relationnel multilingue de noms propres, *Traitement automatique des langues*, Vol. 47(3):115-139[8].

VISEUR R. (2013). Wikipedia. Proceedings of *International Conference on Data Technologies and Applications*. Islande.

YU A. Z., et al. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data* 2:150075.doi: 10.1038/sdata.2015.75.

[8] http://www.atala.org/-Varia,47-