Découverte d'indicateurs de classement à partir de très grands graphes de connaissances

Hassan Abdallah, Béatrice Markhoff, Louise Parkin, Arnaud Soulet

Découverte d'indicateurs de classement à partir de très grands graphes de connaissances

Hassan Abdallah*[0009-0000-5119-6523], Béatrice Markhof**[0000-0002-5171-8499]
Louise Parkin*[0000-0002-6522-3526], Arnaud Soulet*[0000-0001-8335-6069]

*LIFAT, University of Tours
3 Pl. Jean Jaurès, 41000 BLOIS, France
**UMR 7324 CITERES
CNRS and University of Tours

Résumé. Les indicateurs de classement sont des outils essentiels pour situer des entités les unes par rapport aux autres. S'ils sont largement utilisés en infométrie, beaucoup d'autres domaines n'en disposent pas. Dans cet article, nous présentons une méthode pour découvrir automatiquement des indicateurs de classement à partir de très grands graphes de connaissances. Nous formalisons le problème et proposons un algorithme qui trouve des indicateurs pertinents pour un domaine donné en s'appuyant sur deux critères : la proportion d'entités couvertes par un indicateur et une mesure d'inégalité, le coefficient de Gini. Son implémentation est rendue possible grâce à une approximation du coefficient de Gini, dont l'efficacité est démontrée par nos expériences. Ces dernières confirment aussi que l'algorithme extrait des indicateurs variés et faciles à comprendre. Ce travail a fait l'objet d'une publication à la conférence VLDB2025 (Abdallah et al., 2024).

1 Introduction

Les indicateurs de classement permettent de comparer des entités en comptant des propriétés pertinentes pour leur domaine. Cependant, dans de nombreux domaines, il n'existe pas d'indicateurs de classement explicables pouvant être systématiquement appliqués. Par exemple, si l'on considère le domaine de la peinture, des classements de peintres issus de plusieurs sites web sont présentés dans la table 1. Ces classements, tous différents, démontrent non seulement l'intérêt de proposer différents indicateurs de classement, mais aussi la nécessité de fournir une explication pour chacun d'eux.

Les travaux existants concernant le classement d'entités se focalisent sur l'étude d'indicateurs de classement prédéfinis. Gale et Marian (2020) établissent des critères que devraient respecter les indicateurs de classement : interprétabilité, robustesse à des tentatives de manipulations, équité. Dans ce travail, nous proposons une méthode pour identifier automatiquement des indicateurs de classement transparents, explicables et diversifiés, en utilisant les grands graphes de connaissances ouverts du Web.

Les graphes de connaissances ouverts comme Wikidata (Vrandečić et Krötzsch, 2014) contiennent les grand volumes de données nécessaires à la construction de tels indicateurs

(a) theartwolf.com

#	Painter
1	**Picasso**
2	di Bondone
3	**da Vinci**
4	Cézanne
5	**Rembrandt**
6	Velázquez
7	Kandinsky
8	**Monet**
9	Caravaggio
10	van Eyck

(b) ranker.com

#	Painter
1	**Rembrandt**
2	van Gogh
3	Caravaggio
4	Michelangelo
5	**da Vinci**
6	**Monet**
7	Vermeer
8	Raphael
9	**Picasso**
10	Velázquez

(c) artcyclopedia.com

#	Painter
1	**Picasso**
2	van Gogh
3	**da Vinci**
4	**Monet**
5	Dali
6	Matisse
7	**Rembrandt**
8	Warhol
9	O'Keeffe
10	Michelangelo

TAB. 1 – *Trois classements de peintres venant de trois sites Web*

et la sémantique associée à ces données permet de créer des indicateurs pertinents. Pour autant, la découverte automatique d'indicateurs de classement de qualité est complexe : Wikidata est une encyclopédie générale or les indicateurs doivent être spécifiques à un domaine choisi. Ils doivent aussi être calculables à partir de points d'accès SPARQL, en respectant leurs politiques d'usage. Notre approche s'inspire des travaux en bibliométrie (Pratt, 1977) et en altmétrie (Bornmann, 2014) qui ont historiquement utilisé des techniques de comptage pour évaluer l'impact de publications scientifiques. Ces techniques se sont également étendues au domaine du web, avec la naissance de la webométrie (Thelwall, 2008; Stuart, 2014). Toutefois, cette dernière ne tient pas compte de la diversité sémantique des liens que l'on trouve dans des graphes de connaissances. C'est sur ces relations que repose la solution que nous présentons.

Ce travail a fait l'objet d'une publication à la conférence VLDB 2025 (Abdallah et al., 2024) et les codes sources et les résultats sont publics [1]. Nos contributions sont les suivantes :

— Une formalisation des *indicateurs de classement* sous la forme de motif de graphe.
— Une métrique de pertinence de l'indicateur basée sur sa couverture et son degré d'inégalité, ce dernier étant représenté par une formule approchée du coefficient de Gini.
— Un algorithme de recherche efficace, RIPM, qui fournit les indicateurs à la demande en respectant les contraintes d'un point d'accès SPARQL.
— Une évaluation expérimentale sur Wikidata qui montre la précision de l'approximation du coefficient de Gini et une étude utilisateur validant la pertinence et l'interprétabilité des indicateurs de classement générés.

Nous introduisons les notations nécessaires dans la section 2. Nous définissons les indicateurs de classement générés par notre méthode dans la section 3, avant de montrer comment sélectionner les plus pertinents dans la section 4. Notre approche est évaluée en section 5.

2 Préliminaires

Nous nous appuyons sur les notations de Pérez et al. (2009); Hogan et al. (2021).

Graphe de connaissances Soient I et L deux ensembles infinis et disjoints d'IRIs et de littéraux. Un graphe de connaissances est un ensemble de triplets $\mathcal{K} \subseteq I \times I \times (I \cup L)$, où un

1. https://scm.univ-tours.fr/habdallah/RIPM/

fait $\langle s, p, o\rangle \in \mathcal{K}$ se compose d'un sujet $s \in I$, un prédicat $p \in I$, et un objet $o \in I \cup L$ (par exemple, $\langle \texttt{Guernica}, \texttt{creator}, \texttt{Picasso}\rangle$).

Motif de graphe Soit V un ensemble infini de variables. Un *motif de triplet* est de la forme $t \in (I \cup V) \times (I \cup V) \times (I \cup V \cup L)$, où des variables peuvent apparaître dans les trois positions (par exemple, $\langle \texttt{?item}, \texttt{creator}, \texttt{M.G. Benoist}\rangle$). Un *motif de graphe* ou *basic graph pattern* (BGP) est une conjonction de motifs de triplets : si P_1 et P_2 sont des motifs de triplet ou des BGP alors $(P_1 \wedge P_2)$ est un BGP. On note $\mathrm{var}(P)$ les variables du motif P, et $P(?v, i)$ le remplacement de $?v$ par $i \in I$ dans P.

Évaluation Un *mapping* (partiel) est une fonction $\mu : V \to (I \cup L)$. Pour un motif de triplet t, on obtient $\mu(t)$ en remplaçant les variables de t d'après μ. Le domaine $\mathrm{dom}(\mu) \subseteq V$ correspond à l'ensemble de définition de μ. Deux mappings μ_1, μ_2 sont *compatibles* si pour tout $?x \in \mathrm{dom}(\mu_1) \cap \mathrm{dom}(\mu_2)$ $\mu_1(?x) = \mu_2(?x)$. L'évaluation d'un BGP P sur $\mathcal{K}$, noté $[\![P]\!]_{\mathcal{K}}$, est défini récursivement : si P est un motif de triplet t, alors $[\![P]\!]_{\mathcal{K}} = \{\ \mu \mid \mathrm{dom}(\mu) = \mathrm{var}(t),\ \mu(t) \in \mathcal{K}\ \}$. Si $P = (P_1 \wedge P_2)$, alors $[\![P]\!]_{\mathcal{K}} = [\![P_1]\!]_{\mathcal{K}} \bowtie [\![P_2]\!]_{\mathcal{K}}$. Enfin, pour $?v \in \mathrm{var}(P)$ on note $val(P, ?v) = \{\ \mu(?v) \mid \mu \in [\![P]\!]_{\mathcal{K}}\ \}$ et $val_d(P, ?v)$ la version sans doublons de $val(P, ?v)$.

3 Indicateurs de classement

Un indicateur de classement correspond à un comptage pour une entité de faits qui la caractérisent. Dans le contexte d'un graphe de connaissances, on peut préciser ces faits en comptant uniquement certaines propriétés. Par exemple pour un peintre, on peut s'intéresser à ses créations avec la propriété *creator*. Pour affiner davantage on peut restreindre les objets de cette propriété à un certain type, par exemple compter uniquement les peintures, non pas les sculptures ou autres créations du peintre. Nous formalisons un tel indicateur de classement comme un motif de graphe, illustré dans la figure 1 par les deux triplets $\langle ?p, \texttt{creator}, ?ent\rangle$ et $\langle ?p, \texttt{instance of}, \texttt{painting}\rangle$.

Plus précisément, nous commençons par spécifier les entités à classer, toujours grâce à un motif de graphe, le *range pattern* ou motif cible du classement. Un motif cible P est un motif de graphe avec une unique variable $?ent$, dont les valeurs à l'évaluation correspondront aux entités à classer. Dans la figure 1, le motif cible P est $\langle ?ent, \texttt{occupation}, \texttt{painter}\rangle$. D'autres filtres peuvent être ajoutés sur la variable $?ent$, par exemple en utilisant les propriétés *instance of*, *citizenship*, *time period*, etc.

Nous définissons les indicateurs de classement pour les instances du motif cible P comme étant des motifs de graphe de comptage (counting graph patterns). Un motif de graphe de comptage est un motif de chemin acyclique, dont chaque noeud peut être filtré par des propriétés.

Définition 1 (Counting path pattern, ou motif de chemin acyclique)**.** *Un counting path pattern CP est une conjonction de motifs de triplets acyclique $\{\langle ?v_1, p_1, ?v_2\rangle \wedge \langle ?v_2, p_2, ?v_3\rangle \wedge \cdots \wedge \langle ?v_{n-1}, p_n, ?v_n\rangle\}$ telle que $?v_i = ?v_j \Rightarrow i = j$ for $i, j \in \{1, \ldots, n\}$.*

Dans la figure 1, $CP_{sp} = \langle ?item, \texttt{main subject}, ?p\rangle \wedge \langle ?p, \texttt{creator}, ?ent\rangle$ identifie les entités dont le sujet principal est une création de l'entité à classer. $?item$ et $?ent$ sont les variables qui apparaissent respectivement uniquement comme sujet et uniquement comme objet dans CP. Enfin, $?ent$ doit être l'unique variable du motif cible P.

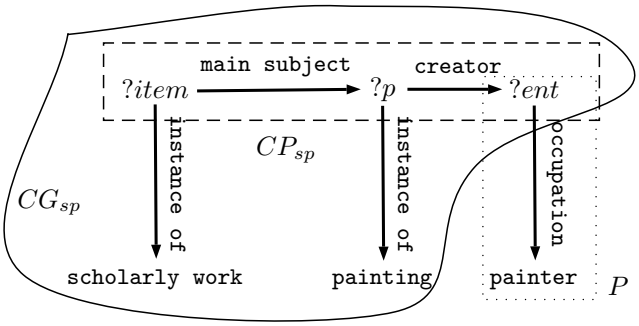

FIG. 1 – *Le counting graph pattern CG_{sp} avec son counting path pattern CP_{sp} et un motif cible P*

Définition 2 (Counting graph pattern, ou motif de graphe de comptage)**.** *Un counting graph pattern CG est un motif de graphe dans lequel il existe un unique counting path pattern $CP \subseteq CG$ et où $var(CP) = var(CG)$. Ce chemin unique est noté $\widetilde{CG}$.*

Intuitivement, les triplets de $CG \setminus \widetilde{CG}$ sont des restrictions sur les variables de $\widetilde{CG}$ de la forme $\langle ?v, p, o\rangle$ ou $\langle s, p, ?v\rangle$ avec $?v \in var(\widetilde{CG})$, $s, p \in I$ et $o \in I \cup L$. Cela permet de raffiner CP_{sp} pour cibler les articles qui concernent une peinture : $CG_{sp} = CP_{sp} \wedge \langle ?item, \texttt{instance of}, \texttt{scholarly work}\rangle \wedge \langle ?p, \texttt{instance of}, \texttt{painting}\rangle$ (cf. figure 1). Nous nommons indicateur de classement (ou *ranking indicator (RI)*) la conjonction d'un counting graph pattern avec le range pattern. Pour un indicateur de classement, nous pouvons associer un score à chaque entité (e.g., compter le nombre de peintures pour chaque peintre) et donc, obtenir le classement des entités selon cet indicateur.

Définition 3 (Comptage pour une entité)**.** *Pour $e \in I$ et $RI \in \mathcal{RI}$, où $\mathcal{RI}$ est l'ensemble des indicateurs de classement, on note $\#RI(e) = \big|val\big(RI(?ent, e), ?item\big)\big|$.*

Dans l'exemple de la figure 1 exécuté sur Wikidata, en utilisant le motif de graphe de comptage $CG_p = \langle ?item, \texttt{creator}, ?ent\rangle \wedge \langle ?item, \texttt{instance of}, \texttt{painting}\rangle$, on obtient $\#RI_p(\texttt{Picasso}) = 818$ (nombre de peintures créées par Picasso) et avec CG_{sp} on obtient $\#RI_{sp}(\texttt{Picasso}) = 1$ (articles concernant les peintures de Picasso). Le classement induit par RI sur $val_d(P, ?ent)$ est l'ordre partiel $e_1 \leq_{RI} e_2 \iff \#RI(e_1) \geq \#RI(e_2)$.

Tous les indicateurs de classement ne sont pas pertinents, c'est pourquoi nous montrons comment les choisir dans la section suivante.

4 Évaluation efficace de la pertinence

S'il existe dans la littérature de nombreuses propositions de classements d'entités, il n'existe, à notre connaissance, aucune caractérisation de ce qui fait un bon indicateur de classement. Intuitivement, on peut considérer que les indicateurs les plus intéressants sont ceux qui distinguent clairement les entités à ranger en produisant un ordre total. Ainsi, nous décomposons la pertinence en deux mesures : la *proportion* des entités couvertes et le *coefficient de Gini* qui capture la dispersion entre les entités.

Proportion Un classement est d'autant plus utile qu'il intègre un grand nombre d'entités du domaine ciblé par P, puisque davantage d'entités peuvent être comparées. Par exemple, plus de peintres sont concernés par le fait d'avoir créé des peintures que par le fait d'avoir créé des sculptures.

Définition 4 (Proportion)**.** *Soit un indicateur de classement RI avec son range pattern P, la proportion des entités couvertes est : $Prop(RI) = \frac{n_e}{n_P}$ avec $n_e = |val_d(RI, ?ent)|$, $n_P = |val_d(P, ?ent)|$.*

Comme $val_d(RI, ?ent) \subseteq val_d(P, ?ent)$, $Prop(RI) \in [0, 1]$. Pour les peintres dans Wikidata, en considérant leur nombre de peintures la proportion de peintres concernés est $Prop = 61\% = 178\,126/291\,617$, tandis qu'en comptant leurs sculptures, elle est seulement de $Prop = 7\% = 19\,576/291\,617$. Le premier critère conduit donc à un meilleur classement, qui permet de comparer plus de peintres. L'augmentation de la couverture augmente le nombre de paires strictement comparables, et tend vers 1 pour un indicateur de classement idéal.

La couverture n'est pourtant pas suffisante pour garantir un classement de qualité, car il est aussi nécessaire d'avoir une bonne dispersion. Pour cela, nous utilisons le coefficient de Gini.

Coefficient de Gini Le coefficient de Gini est une mesure de concentration très utilisée en économie et en bibliométrie (Pratt, 1977).

Définition 5 (Coefficient de Gini)**.** *Pour un ensemble d'entités $\langle e_1, e_2, \ldots, e_{n_e} \rangle$ ordonnées selon leur score $\#RI$, le coefficient de Gini est défini par :*

$$Gini(RI) = \frac{2\sum_{i=1}^{n_e}(n_e - i + 1)\,\#RI(e_i)}{n_e \sum_{i=1}^{n_e} \#RI(e_i)} - \frac{n_e + 1}{n_e}.$$

$Gini(RI) \in [0, 1]$, où 0 correspond à une égalité parfaite et 1 à une inégalité maximale.

Le calcul du coefficient de Gini selon cette formule est très coûteux dans le cas de grands range patterns, car il nécessite des comptages individuels par entité et il doit être répété sur de nombreux indicateurs candidats. Nous proposons de calculer plutôt une approximation de sa valeur, en utilisant le modèle des relations proposé par Abdallah et al. (2025), où la distribution des faits sur les entités suit une loi de puissance d'exposant $\alpha = 1 + \frac{1}{1-n_e/n}$.

Théorème 1 (Approximation du Gini)**.** *Pour un indicateur de classement RI, le Gini peut être approximé par*

$$\widetilde{Gini}(RI) = \frac{1 - n_e/n}{1 + n_e/n},$$

où $n_e = |val_d(RI, ?ent)|$ et $n = |val(RI, ?item)|$.

Algorithme Nous pouvons donc déterminer la qualité des indicateurs de classement en utilisant seulement trois valeurs : le nombre d'entités à classer $n_P = |val_d(P, ?ent)|$, le nombre d'entités avec un score non nul $n_e = |val_d(RI, ?ent)|$, le nombre total de faits comptés $n = |val(RI, ?item)|$. Pour un graphe de connaissances $\mathcal{K}$ et un range pattern P, nous proposons un algorithme permettant de retourner des indicateurs de classement RI pour les entités ciblées par P qui maximisent la pertinence : $\arg\max_{RI \in \mathcal{RI}} Gini(RI) \times Prop(RI)$.

L'algorithme prend en entrée un graphe de connaissance et un range pattern. Il parcourt ensuite la liste de toutes les propriétés compatibles avec les entités ciblées par le range pattern.

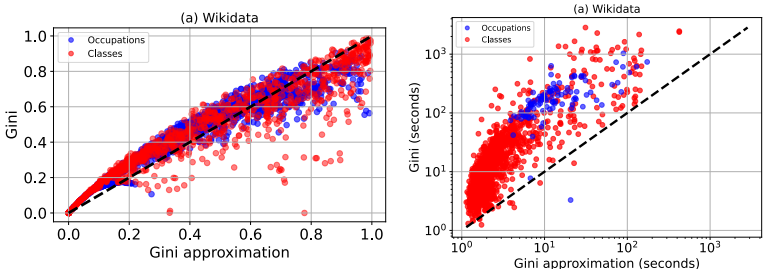

FIG. 2 – *Gauche : $Gini$ versus $\widetilde{Gini}$ pour Wikidata ; droite : temps de calcul pour Wikidata*

Cela permet de générer ensuite tous les CP de longueur 1. Seulement ceux dont le Gini dépasse un seuil γ paramétrable sont conservés. Chacun de ces CP est ensuite étendu en CG avec un unique triplet qui restreint la variable à une instance d'une certaine classe. Les k classes les plus fréquentes sont utilisées pour cette restriction, où k est également un seuil paramétrable. Les scores de Gini $\widetilde{Gini}$ et de proportion $Prop$ pour chaque indicateur ainsi formé peuvent alors être calculés grâce à n_P, n_e, et n. L'algorithme retourne l'ensemble des indicateurs de classement ainsi formés avec leurs scores de Gini et de proportion.

5 Étude expérimentale

Nous avons évalué la précision et l'efficacité de notre approximation du coefficient de Gini, ainsi que l'intelligibilité des indicateurs et leur pertinence pour classer les entités des domaines choisis grâce à une étude avec des utilisateurs.

Protocole Nous avons utilisé les points d'accès public de Wikidata, DBpedia, et YAGO[2] en juin 2024. Les sources sont disponibles sur Zenodo `https://zenodo.org/records/14181263`. Nous avons appliqué RIPM avec $k = 5$ et $\gamma = 0.1$. Les domaines ciblés pour les classements sont les 100 occupations les plus fréquentes, et les classes de sujets qui ont été découvertes en construisant les indicateurs de classement pour les occupations.

Évaluation de l'approximation du coefficient de Gini Nous comparons la valeur exacte $Gini$ avec l'approximation $\widetilde{Gini}(n_e, n) = \frac{1-n_e/n}{1+n_e/n}$. La figure 2 à gauche trace $Gini$ en fonction de $\widetilde{Gini}$ pour les expériences sur Wikidata. On observe un tracé proche de l'identité. Les scores d'erreur sont très faibles avec $\text{MAE} = \frac{1}{n}\sum_i |Gini_i - \widetilde{Gini_i}| = 0,068$ et $\text{MSE} = \frac{1}{n}\sum_i (Gini_i - \widetilde{Gini_i})^2 = 0,008$. La figure 2 représente à droite les temps de calcul pour $Gini$ et $\widetilde{Gini}$. L'approximation permet un gain de 88.9%. Ces observations se vérifient sur l'ensemble des domaines, et sont également valables pour les expériences avec DBpedia et YAGO.

Évaluation de la pertinence par une étude utilisateur

Pour évaluer si notre mesure de pertinence $\widetilde{Gini} \times Prop$ correspond à ce qui est jugé intéressant par des utilisateurs, pour chaque occupation nous avons présenté aux utilisateurs trois indicateurs exprimés en langage naturel extraits des sorties de RIPM : l'indicateur identifié

2. query.wikidata.org, dbpedia.org/sparql, yago-knowledge.org/sparql/query

TAB. 2 – *Accord entre annotateurs et comparaison avec RIPM, Kendall's τ*

Occupation	**relation**	**classe**	**τ entre utilisateurs**	**τ utilisateurs / RIPM**
Writers	author	version, edition or translation	0.610	1.000
Univ. teachers	doctoral advisor	human	0.454	1.000
Singers	performer	album	0.236	1.000
Journalists	author	version, edition or translation	0.252	1.000
Poets	author	version, edition or translation	0.555	1.000
Actors	cast member	film	0.587	1.000
Painters	creator	painting	0.360	1.000
Composers	composer	musical work, composition	0.454	0.333
		Moyenne	**0.438**	**0.917**

comme le meilleurs, un moyen (situé entre le 45eme et 55eme percentile de la liste ordonnée) et un mauvais (tiré dans les derniers 10%). Par exemple pour le classement des peintres, notre protocole propose le nombre de peintures qu'ils ont créées (meilleur), le nombre d'imprimés issus de leurs travaux (moyen), et le nombre de musées qui portent leur nom (mauvais, pénalisé par sa faible couverture). 19 utilisateurs avec une formation en informatique ont ordonné les trois options selon leur pertinence pour l'occupation ciblée. Nous calculons le τ de Kendall entre utilisateurs par occupation, déterminons un classement majoritaire comme vérité de terrain, et le comparons avec le classement produit par RIPM grâce au τ de Kendall. Les résultats sont présentés dans la table 2, ainsi que la relation et la classe correspondant au meilleur indicateur de classement d'après notre mesure. L'accord entre annotateurs est moyen, et l'accord entre le classement majoritaire et RIPM est très élevé, parfait pour 7 occupations sur 8. Cela montre l'intérêt pratique du score $\widetilde{Gini} \times Prop$ pour indiquer la pertinence d'un indicateur de classement, tout en étant rapide à calculer.

6 Conclusion

Nous avons proposé une méthode pour la découverte automatique d'indicateurs de classement à partir de graphes de connaissances. Nos indicateurs sont définis comme des motifs de graphes de comptage qui ont à la fois une grande couverture du domaine ciblé produisent des scores variés entre les entités à classer, fournissant ainsi des classements pertinents. Nous avons décrit un algorithme efficace qui interroge directement un point d'accès SPARQL public. Pour cela, nous avons utilisé une approximation du coefficient de Gini permettant un élagage de l'espace de recherche. Nos expériences sur Wikidata ont montré que cette approximation a une bonne précision et un coût raisonnable. Une étude des classements produits montre leur diversité, et une étude utilisateur montre la forte compatibilité entre les classements identifiés comme les plus intéressants par notre méthode et par les utilisateurs. Ainsi, nous avons montré que les graphes de connaissances permettent de construire des indicateurs de classement explicables et transparents pour des domaines qui n'ont pas d'indicateurs établis. Les perspectives de ce travail consistent en la construction d'indicateurs composites permettant de refléter de multiples critères.

Références

Abdallah, H., B. Markhoff, et A. Soulet (2024). Ranking Indicator Discovery from Very Large Knowledge Graphs. *Proceedings of the VLDB Endowment 18*(4), 1183–1195.

Abdallah, H., B. Markhoff, et A. Soulet (2025). A complex network model for knowledge graphs' relationships. *Semantic Web 16*(5).

Bornmann, L. (2014). Do altmetrics point to the broader impact of research ? an overview of benefits and disadvantages of altmetrics. *Journal of informetrics 8*(4), 895–903.

Gale, A. et A. Marian (2020). Explaining monotonic ranking functions. *Proceedings of the VLDB Endowment 14*(4), 640–652.

Hogan, A., E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al. (2021). Knowledge graphs. *ACM Computing Surveys (Csur) 54*(4), 1–37.

Pérez, J., M. Arenas, et C. Gutierrez (2009). Semantics and complexity of sparql. *ACM Transactions on Database Systems (TODS) 34*(3), 1–45.

Pratt, A. D. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for information Science 28*(5), 285–292.

Stuart, D. (2014). *Web metrics for library and information professionals*. Facet Publishing.

Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of information science 34*(4), 605–621.

Vrandečić, D. et M. Krötzsch (2014). Wikidata : a free collaborative knowledgebase. *Communications of the ACM 57*(10), 78–85.

Summary

Ranking indicators are essential tools for comparing the importance of various entities such as cities or scientists. While extensively used in fields like econometrics and scientometrics, many other domains lack systematic approaches for developing these indicators. In this paper, we introduce a novel method for automatically discovering ranking indicators from very large knowledge graphs. To this end, we formalize the notion of *counting graph pattern* (CG) as a special SPARQL query, and the concept of *ideal ranking indicator* as a CG whose result induces a strict total order on a set of entities. To assess the interestingness of ranking indicators, we employ the *proportion of covered entities* along with an inequality measure, namely the Gini coefficient. We further present Algorithm *Ranking Indicator Pattern Miner (*RIPM*)*, to efficiently identify interesting ranking indicators for a given field. Our experimental study shows the effectiveness of our Gini approximation. It also validates that RIPM extracts *transparent, diverse, and understandable* indicators through a user survey. This work has significant implications for fields lacking dedicated communities working on ranking tasks, providing a robust tool to automatically produce ranking indicators, and the associated rankings.