Revisão Sistemática da Literatura sobre ranking de Relacionamentos na Web Semântica

Alternative Title: A Systematic Literature Review on Relationship Ranking in the Semantic Web

Paulo M. F. dos Santos
Universidade de São Paulo
paulofranco@usp.br

Karina V. Delgado
Universidade de São Paulo
kvd@usp.br

Marcelo de S. Lauretto
Universidade de São Paulo
marcelolauretto@usp.br

Marcio M. Ribeiro
Universidade de São Paulo
marciomr@usp.br

RESUMO

O ato de realizar pesquisas na *Web* tem sido o mesmo por anos. O usuário realiza uma consulta composta de termos, e o motor de busca é responsável por encontrar as melhores respostas àquela consulta. Frequentemente, existem informações subjetivas que o usuário não consegue transmitir em sua consulta, mas que ele espera que o motor de busca seja capaz de inferir. Isso leva a resultados que são relacionados à sua consulta, mas não aos seus interesses. Uma forma de mitigar esse problema foi a introdução da *Web* Semântica, que visa a permitir que os dados disponíveis na *Web* tenham um sentido, ou seja, uma semântica. Diversas abordagens de busca na *Web* Semântica têm sido propostas e implementadas nos últimos anos, bem como abordagens para classificação (*ranking*) de resultados. Esta revisão sistemática da literatura tem por objetivo identificar as tendências na área de *ranking* de relacionamentos na *Web* Semântica. De um total de 1.194 artigos inicialmente retornados em nossa pesquisa, foram selecionados e analisados 10 estudos primários nesse tipo de *ranking*, dando-se especial atenção às características das técnicas adotadas e aos experimentos realizados. Observou-se então que novas soluções promissoras envolvem o uso de algoritmos de aprendizado de máquina para realizar o *ranking* dos resultados das consultas.

Palavras-Chave

Web semântica, Linked data, ranking de relacionamentos.

ABSTRACT

The act of searching the web has been the same for years. The user inputs a query and the search engine is responsible for finding the best matches to that query. Often, there are subjective information that the user can't transmit when making his query, but he expects that the search engine will infer. This leads to results that are query-related, but not user-interest related. One way of mitigating this problem was the introduction of the Semantic Web, wich aims to allow that the data available on the web have a meaning. Many approaches on semantic web search that crawl the semantic web have been proposed and implemented, as well as solutions to better rank and classify the results. This systematic review of the literature aims to obtain knowledge about the latest trends about the semantic relationships on the semantic web. Of a total of 1194 papers initially obtained during the research, 10 were selected and studied on this subject, giving a special atention to the techniques used and the experiments made. It was then observed that promising new solutions involves the use of machine learning algorithms as a means to rank the results of a query.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•**Information systems → Learning to rank;** *Web data description languages;*

Keywords

Semantic *web*, Linked data, Relationship ranking.

1. INTRODUÇÃO

Com a popularização da internet nos anos 90, a quantidade de páginas disponíveis na *World Wide Web* (abreviada informalmente para *Web*) cresceu em um nível exponencial. Cada página da *Web* representa uma fonte de conhecimento e interesse. Essas páginas e seus conteúdos vêm sendo criados e armazenados da mesma forma ao longo dos anos. Motores de busca, que realizam pesquisa sobre a *Web*, vêm se tornando cada vez mais poderosos, com seus algoritmos otimizados para recuperar a maior quantidade de informação, na menor quantidade de tempo possível. No entanto, devido à dificuldade ou impossibilidade de se introduzir informações subjetivas nas consultas, pode ser que uma grande quantidade dos resultados encontrados pelos motores de busca tenham baixa relevância para o usuário [16]. Esse fenômeno é observado em motores de busca baseados exclusivamente em consultas por palavras-chave.

Para mitigar esse problema, foi criada uma proposta de evolução da *Web*, denominada *Web Semântica*. A *Web* Semântica é uma extensão da *Web* que introduz padrões para o compartilhamento de dados na rede, de forma a permitir que os dados disponíveis passem a ter um sentido (uma semântica). Os dados podem assim ser distinguidos pelas máquinas, que agora trabalham com o próprio conhecimento acerca das informações, ao invés de trabalhar apenas com textos, facilitando a recuperação automática de informações.

A área de pesquisa em *Web* Semântica tem evoluído significativamente, com o objetivo de mudar a maneira em que as informações são organizadas, armazenadas e recuperadas na *Web*. Motores de busca semântica têm sido desenvolvidos na última década, com o foco de recuperar informações que sejam mais relevantes para o usuário. Mais recentemente, os dados na *Web* Semântica mudaram da maneira caótica em que eram organizados e distribuídos para uma maneira mais fácil e concisa, denominada *Linked Data*. *Linked Data* são uma série de boas práticas desenvolvidas para organizar os dados na *Web* em entidades e relacionamentos, interconectando-os em estruturas de grafos [5]. Atualmente, a organização e disponibilização desses dados se dá através do *Resource Description Framework* (RDF), um modelo de organização de dados que se aproxima de conceitos de modelagem clássicos de bancos de dados.

Apesar de existir revisões sobre *ranking* na *Web* Semântica [12, 28, 20], em buscas preliminares não encontramos revisões sistemáticas (RSs) da literatura nesse tema. Adicionalmente, as revisões conhecidas apresentam trabalhos em períodos mais antigos, de até duas décadas.

O objetivo desta Revisão Sistemática (RS) é analisar os principais artigos publicados desde 2005 que contemplem técnicas para *ranking* de relacionamentos na *Web* Semântica, com especial enfoque nas características das técnicas utilizadas e nos experimentos realizados.

O restante deste texto está organizado da seguinte forma. A Seção 2 apresenta os conceitos fundamentais incluindo a definição de *ranking*, a classificação dos algoritmos de *ranking* para a *Web* Semântica e a definição de associação semântica. A Seção 3 apresenta o método utilizado e a Seção 4 apresenta uma discussão dos resultados obtidos. Por fim, a Seção 5 apresenta os comentários finais e conclusões.

2. RANKING DE DADOS NA WEB SEMÂNTICA

Um dos fatores importantes na recuperação de informação é como organizar e apresentar os resultados das buscas. Para tal, é necessário escolher uma classificação subjetiva de importância para esses dados. Essa organização dos resultados através de um critério subjetivo é denominada *ranking*. No caso da Web Semântica, existem diversos critérios que podem ser adotados para formar diferentes *ranking*s das informações que a constituem. É possível organizar e classificar seus dados através da importância relativa das propriedades RDF, ou através de critérios como popularidade das entidades, raridade dos conceitos, entre outros.

2.1 Classificação das abordagens de ranking para a Web Semântica

Jindal et al. [12] classificam as abordagens de *ranking* para a Web Semântica em três categorias: *ranking* de entidades, *ranking* de relacionamentos e *ranking* de documentos.

Os algoritmos de *ranking* de entidade são aqueles que têm como objetivo recuperar resultados baseados nas entidades de interesse fornecidas como entrada pelos usuários ao motor de busca. Esses algoritmos encontram a proximidade da entidade com as entidades vizinhas baseados na quantidade de relacionamentos entre elas [12]. Alguns exemplos de estudos que fazem esse tipo de *ranking* são [26, 19, 11].

Os algoritmos de *ranking* de relacionamento são aqueles que focam na importância relativa dos relacionamentos entre as entidades, com relação ao contexto da consulta do usuário. Nesta RS, estamos interessados nesta abordagem de *ranking* pois os relacionamentos são a parte mais importante da semântica dando significado à informação, tornando-a compreensível e fornecendo conhecimentos novos e possivelmente inesperados sobre a informação [1].

Os algoritmos de *ranking* de documentos são os que buscam documentos com o mais relevante e completo conjunto de entidades de interesse, bem como o conjunto mais relevante de relacionamentos entre essas entidades [12]. Entre os trabalhos que seguem essa abordagem estão [1, 17, 7].

2.2 Busca de Associações Semânticas e ranking

Seja o grafo $G = (E, P)$, em que E é o conjunto de vértices que representam as entidades e P é o conjunto de arestas que representam os relacionamentos ou também chamadas de propriedades entre as entidades. Duas entidades e_1 e e_n estão semanticamente associadas se existe uma sequencia $\langle e_1, p_1, e_2, p_2, \cdots e_{n-1},\ p_{n-1}, e_n \rangle$ em que $e_i \in E$, $p_i \in P$ e $1 \leqslant i \leqslant n$. Em outras palavras, associações semânticas são definidas de forma geral como o conjunto de sequências diferentes de propriedades que conectam duas entidades no modelo RDF [1].

Exemplo 1: Entre Brasil e Paraguai, existem diversas associações semânticas, sendo a mais direta e explícita que existe fronteira entre os dois países. Outra associação possível é que *Brasil* seja um *País*, que é membro do *Mercosul*, que por sua vez é um *Bloco Econômico*, do qual o *Paraguai* também faz parte. Essa é uma associação mais extensa entre eles, e não explícita, já que o caminho que os conecta passa por diversas outras entidades.

A busca de associações semânticas recebe um par de entidades $(e1, e2)$ como entradas da consulta, e devolve o conjunto de todos os caminhos no grafo que interconectam essas duas entidades, resultando assim em uma coleção de associações semânticas não-ordenadas.

Existem diversas linguagens criadas para fazer *query* de associações semânticas, como SPARQLeR [15] e *ρ-queries* [4]. Porém elas não permitem ao usuário incluir suas preferências de maneira explícita. Vários esforços de pesquisa têm sido realizados para identificar relações relevantes a partir dos resultados retornados pela *query* de associações semânticas. Entre os trabalhos citados em [12] que fazem esse tipo de *ranking* estão *SemRank* [3] e *SemDis* [2].

3. MÉTODO

Uma revisão sistemática da literatura é um estudo secundário que fornece um processo rigoroso e replicável para rever as evidências relevantes para uma determinada questão de pesquisa. Em [14] são apresentas as três etapas principais de uma RS: planejamento, condução e relatório. A etapa de planejamento produz um protocolo que descreve os componentes da revisão sistemática e permite a replicabilidade do

método. A etapa de condução consulta o protocolo para selecionar os estudos e fornecer os dados para responder as perguntas específicas de pesquisa. Na etapa de relatório, esses dados são divulgados e analisados em um documento.

3.1 Planejamento

Questões de pesquisa. As questões de pesquisa são:

Q-1 Que técnicas são utilizadas para fazer *ranking* de relacionamentos na *Web* Semântica?

A questão Q-1 visa identificar a tendência das técnicas utilizadas para fazer *ranking* de relacionamentos na *Web* Semântica.

Q-2 Quais as características da abordagem utilizada?

- Q-2.1 A abordagem adotada é independente do domínio ou de um domínio específico?
- Q-2.2 A abordagem adotada é independente da *query* ou de uma *query* específica?
- Q-2.3 Qual a interface da *query*?
- Q-2.4 O sistema é capaz de aprender a preferência do usuário com o passar do tempo?
- Q-2.5 O sistema cria uma função de *ranking* personalizada para cada usuário?
- Q-2.6 O usuário configura manualmente os parâmetros de *ranking*?
- Q-2.7 A abordagem é sensível ao contexto?

O objetivo de Q-2 é identificar as características da abordagem utilizada, para isso, Q-2 foi decomposta em questões mais especificas. A abordagem é dependente de domínio se ela é limitada a uma determinado conjunto de dados que abordam um assunto específico, como por exemplo uma ontologia sobre filmes de ficção; e independente de domínio se ela não possui essa limitação. A abordagem adotada é dependente da *query* se a função de *ranking* implementada é afetada pelo conteúdo que o usuário insere na busca, e independente da *query* se a função de *ranking* se baseia somente na estrutura interna do conjunto de dados. A interface da *query* trata da maneira que o usuário interage com o sistema, seja através do uso de palavras-chave, de linguagem natural, de línguagem estruturada de *query*, ou outras.

Q-3 Quais as características dos experimentos realizados?

- Q-3.1 Qual a base de dados utilizada?
- Q-3.2 Que métricas são usadas para comparar os resultados?
- Q-3.3 Com quais outras técnicas foi comparada a abordagem proposta?
- Q-3.4 Qual a quantidade de instâncias e relações da base de dados utilizada nos experimentos?
- Q-3.5 Foram realizadas interações com humanos nos experimentos de *ranking*?

A questão Q-3 busca identificar como os experimentos foram realizados, as bases de dados utilizadas, as métricas analisadas e se houve comparação com outras técnicas.

Seleção de fontes e string de busca. Foram escolhidas as seguintes fontes: Scopus[1] e *Web* of Science[2]. Note que o motor de busca Scopus indexa outras fontes de dados, entre elas IEEE Xplore[3], Science Direct[4] , ACM Digital Library[5] e Engineering Village[6]. Na Tabela 1 são apresentadas as duas strings de busca usadas nesta RS, uma para cada motor de busca. Nessas strings foram incluídos os termos importantes e seus sinônimos nos campos título, resumo e palavras-chave dos artigos.

Table 1: String de busca usado na RS

(TITLE-ABS-KEY (("semantic Web" OR "semantic relationship" OR "semantic search*" OR "linked data") AND ("rank*" OR "rank* algorithm*")))*
(TS=(("semantic Web" OR "semantic relationship" OR "semantic search*" OR "linked data") AND ("rank*" OR "rank* algorithm*")))*

Critérios de Inclusão e Exclusão

Os critérios de inclusão (CI) e exclusão (CE) usados nesta RS têm como objetivo selecionar estudos primários com um nível mínimo de qualidade e que não fujam do objetivo desta RS. Os critérios aplicados foram:

CI-1 Os artigos devem estar disponíveis de forma integral e gratuita online, ou os autores devem ter acesso aos artigos através da sua instituição.

CE-1 Artigos que não abordem principalmente algoritmos de *ranking* de relacionamentos da *Web* semântica.

CE-2 Artigos publicados por meios que não exigem revisão por pares.

CE-3 Estudos secundários ou terciários.

CE-4 Estudos não relacionados às áreas de Ciência da Computação.

CE-5 Estudos incompletos ou que não relatam de maneira adequada os algoritmos de *ranking*.

CE-6 Estudos não publicados em *workshop*, conferência ou periódico.

CE-7 Artigos que não tenham sido publicados apartir de 2005.

CE-8 Artigos que não estão inteiramente escritos na língua inglesa.

3.2 Condução

Para auxiliar na condução desta RS foi utilizado o Software StArt *(State of the Art through Systematic Review)*[7]. A pesquisa foi realizada no dia 31 de janeiro de 2017. O motor de busca Scopus retornou 1423 trabalhos, dentre os quais 326 artigos foram excluídos pelos critérios CE-2, CE-3, CE-4, CE-6, CE-7 e CE-8 através das ferramentas de filtro

[1]https://www.scopus.com/
[2]http://apps.*Web*ofknowledge.com
[3]http://ieeexplore.ieee.org
[4]http://www.sciencedirect.com
[5]http://dl.acm.org
[6]http://www.engineeringvillage.com
[7]Disponível em: http://lapes.dc.ufscar.br/tools/start_tool

desse motor de busca. No motor do *Web* of Science foram retornados 556 estudos, dentre os quais 27 artigos foram excluídos pelos critérios CE-2, CE-3, CE-6, CE-7 e CE-8 também através das ferramentas de filtro.

Foram identificados 432 artigos duplicados. Assim, do total de 1626 artigos, restaram 1194 estudos para serem analisados. Note que a quantidade de trabalhos restantes é ainda relativamente grande, pois as palavras-chave *semantic Web* e *ranking* são termos bastante usados na literatura. Entre esses trabalhos, além dos trabalhos que fazem *ranking* de relacionamentos, foi encontrado um grande número de estudos que fazem *ranking* de entidades e *ranking* de documentos. Os critérios de inclusão e exclusão foram aplicados manualmente em cada um dos 1194 estudos restantes. Para cada um deles foi analisado o título, resumo e palavras chaves, e no caso de dúvida o estudo foi analisado por completo. No fim dessa etapa foram selecionados 10 estudos para o desenvolvimento desta RS.

4. RESULTADOS

Na Tabela 2 são apresentados os artigos selecionados após a execução do protocolo da revisão. A primeira coluna apresenta o ano de publicação do artigo e a segunda o identificador (ID) em que a letra J refere-se a artigo de periódico e a letra C refere-se a artigo de conferência. A coluna 3 apresenta a referência bibliográfica de cada artigo, e a coluna 4, o título. Os artigos estão ordenados pelo ano de publicação.

Antes de responder às questões de pesquisa, são descritas as diferentes métricas de associações semânticas encontradas em alguns dos estudos selecionados nesta RS.

4.1 Métricas de associações semânticas

Entre as métricas de associações semânticas encontradas estão o comprimento, a popularidade, a raridade, subsunção e confiança, propriedades tópicas, complexidade da relação, frequência das propriedades, ganho de informação, refração e importância com relação às palavras-chave.

Comprimento (L_A): se refere a quantas relações intermediárias existem entre duas entidades e_1 e e_2. No Exemplo 1, a primeira associação tem comprimento um e a segunda comprimento dois.

Popularidade (P_A): dada uma base de dados semântica, algumas associações são mais frequentes que outras. Em uma base de dados de artigos científicos, por exemplo, é mais fácil encontrar relações entre autores que citam trabalhos de outros autores, do que relações de parentesco entre os autores, portanto as relações de citação são mais populares nesse cenário.

Raridade (R_A): algumas vezes, um usuário pode justamente procurar conhecimentos que ele não esperaria encontrar com facilidade, pois são incomuns dado o contexto.

Subsunção (S_A): refere-se à especificação da hierarquia do contexto. Classes mais baixas na hierarquia de classes possuem significado mais específico, e acabam ganhando maior relevância. Por exemplo, a entidade *professor* transmite mais significado do que uma entidade *pessoa*.

Confiança(T_A): refere-se ao quanto o usuário confia na origem dos dados. Como as entidades e relacionamentos em uma associação semântica são de origem diferente, cabe ao usuário especificar o quanto confia na fonte daqueles dados.

Propriedades tópicas (PT_A): que mede a quantos tópicos uma entidade pertence.

Complexidade da relação ($Complex_A$): que mede a complexidade de uma associação semântica baseado na quantidade de propriedades que ela possui.

Frequência das propriedades (F_A): são usadas a média, o desvio padrão, o mínimo e o máximo das frequências das propriedades de uma associação semântica.

Ganho de informação (G_A): o ganho de informação de uma propriedade G_P é a quantidade de incerteza removida ao saber que uma propriedade específica acontece. O ganho de uma sequencia de propriedades G_A é o máximo ganho de todas as propriedades da sequencia.

Refração (Ref_A): a refração mede o desvio do caminho da associação semântica de um esquema de dados para outro. Um caminho com muitas refrações tem menos chance de ser facilmente antecipado pelo usuário, tornando-o mais imprevisível.

Importância com relação às palavras-chave ($KMatch_A$): o usuário insere um conjunto de palavras-chave que estão relacionadas com as propriedades de interesse. Dada uma sequência $\langle e_1, p_1, e_2, p_2, \cdots e_{n-1},\ p_{n-1}, e_n\rangle$ em que $e_i \in E$ e $p_j \in P$, $KMatch_A$ é a soma dos níveis na hierarquia de propriedades em que a palavra-chave coincide com a propriedade p_i.

4.2 Técnicas utilizadas para fazer ranking de relacionamentos: Q-1

4.2.1 Estudos que usam métricas das associações semânticas e não usam aprendizado de máquina

Em **J-1**, os autores introduziram métricas semânticas para fazer o *ranking* de associações semânticas entre duas entidades. Foram introduzidas as métricas de associações semânticas L_A, P_A, R_A, S_A, T_A e contexto.

Para calcular o valor do contexto, denotado por C_A, é exibida uma ontologia para o usuário, usando uma ferramenta gráfica. Nessa ontologia, o usuário define qual região (contexto) é mais importante para ele. Baseadas nessa informação, as associações semânticas ganham pesos de acordo com as regiões em que estão envolvidas. A equação proposta em J-1 para fazer o *ranking* das associações, considerando todas essas métricas, é dada por:

$$W_A = k_1 \times C_A + k_2 \times S_A + k_3 \times T_A + k_4 \times R_A + k_5 \times P_A + k_6 \times L_A, \quad (1)$$

em que $k_i (1 \leqslant i \leqslant 6)$ são os pesos de preferência, que podem ser ajustados pelo usuário, e $\sum k_i = 1$.

Em **C-1**, os autores discutem uma outra forma de incluir o contexto da consulta para realizar o *ranking* de associações semânticas. O conceito de fator de confiança de contexto de cada entidade é introduzido considerando que as entidades podem fazer parte de mais de um contexto. O fator de confiança representa o grau de relevância da entidade para um contexto particular. Esse fator é utilizado para calcular o peso das associações semânticas, chamado de $Context_A$, em um determinado contexto. O *ranking* das associações semânticas em C-1 é calculado pela equação:

$$W_A = k_1 \times L_A + k_2 \times P_A + k_3 \times Context_A, \quad (2)$$

em que $k_i (1 \leqslant i \leqslant 3)$ são os pesos de preferência determinados pelo usuário e $\sum k_i = 1$

Em **C-2**, os autores propõem uma forma de encontrar os k relacionamentos mais relevantes para um usuário. Para realizar o cálculo da relevância das associações, os autores consideram as métricas de associações semânticas G_A, Ref_A,

Table 2: Estudos selecionados nesta RS

Ano	ID*	Referência	Título
2005	J-1	[2]	*ranking* Complex Relationships on the Semantic *Web*
2007	C-1	[13]	A Context-Aware *ranking* Method for the Complex Relationships on the Semantic *Web*
2008	C-2	[21]	An RDF Approach for Discovering the Relevant Semantic Associations in a Social Network
2008	C-3	[27]	Identifying Potentially Important Concepts and Relations in an Ontology
2009	C-4	[18]	Semantic association search and rank method based on spreading activation for the Semantic *Web*
2012	J-2	[9]	Learning to Rank Complex Semantic Relationships
2012	C-5	[8]	Rankbox: An adaptive *ranking* system for mining complex semantic relationships using user feedback
2012	C-6	[25]	*ranking* Semantic Associations between Two Entities – Extended Model
2012	J-3	[24]	*ranking* semantic relationships between two entities using personalization in context specification
2016	C-7	[6]	Separating Wheat from the Chaff – A Relationship *ranking* Algorithm

*J=journal, C=conference

P_A e $KMatch_A$. Essas métricas são combinadas de diferentes formas, dependendo do modo de pesquisa, que pode ser: (i) busca por associações comuns, que são caracterizadas por propriedades de ocorrência frequente, nós altamente conectados e um único domínio ou (ii) busca por associações informativas, que contém as propriedades menos frequentes, nós menos conectados e múltiplos domínios. Para fazer o *ranking* de associações comuns é usada a equação:

$$W_{A_C} = G_A^{-1} + P_A + KMatch_A \quad (3)$$

e para associações informativas a equação é:

$$W_{A_I} = G_A + (1 - P_A) + Ref_A + KMatch_A, \quad (4)$$

Em **C-6**, o algoritmo de *ranking* proposto funciona de forma similar a J-1, com a diferença que os autores adicionaram o conceito de proximidade do contexto. Dada uma sequencia $\langle e_1, p_1, e_2, p_2, \cdots e_{n-1}, p_{n-1}, e_n \rangle$ em que $e_i \in E$ e $p_j \in P$, e_1 e e_n são chamadas de entidade da esquerda e da direita da relação, respectivamente. O contexto é mais próximo da entidade da esquerda se $\sum_{i=1}^{n/2} peso(e_i) > \sum_{n/2+1}^{n} peso(e_i)$, caso contrário o contexto é mais próximo da entidade da direita. Naquele trabalho, o usuário primeiro deve definir o seu contexto de interesse na hora de realizar a consulta para definir os pesos. Em seguida, os fatores de proximidade são calculados para saber de qual das duas entidades, esquerda ou direita, o contexto é mais próximo. Baseado nessa escolha, o valor de contexto MC_A é calculado. De forma geral, a equação de *ranking* de C-6 é similar à Equação 1, substituindo apenas C_A por MC_A.

Em **J-3**, o algoritmo de *ranking* proposto funciona de forma similar a C-6, baseando-se ainda em J-1, com a diferença que os autores adicionaram mais um fator de personalização baseado no histórico de navegação. O primeiro passo do algoritmo é construir uma tabela de nível de interesse do usuário em várias categorias do domínio, através do seu histórico de navegação. O segundo passo é calcular o valor do contexto personalizado para um dado usuário (MC_Mod_A) usando essa tabela. Assim, a especificação do contexto é feita sem a intervenção do usuário, isto é, de forma automatizada. Por último, é feito o *ranking* dos relacionamentos usando uma equação similar à Equação 1 em que C_A e substituída por MC_Mod_A.

4.2.2 Estudos que usam métricas das associações semânticas e aprendizado de máquina

Em **J-2**, os autores utilizam aprendizado de máquina, mais especificamente Support Vector Machines (SVMs), como forma de realizar o *ranking* de associações semânticas complexas. Os autores afirmam que seu algoritmo é capaz de trabalhar com qualquer quantidade de métricas sobre as relações e as entidades, podendo inclusive utilizar as mesmas definidas em J-1, ou em outros trabalhos. Em J-2 é apresentado um exemplo com um conjunto de métricas de associações semânticas, em que as métricas L_A, PT_A, $Complex_A$, F_A e P_A são utilizadas. Com base nisso, monta-se um vetor de características x_A. O usuário é levado a treinar a função de *ranking*, classificando manualmente o resultado de um conjunto de consultas apresentadas a ele. Essas consultas atualizam os valores do vetor de pesos w da função de *ranking*, que é um vetor de mesmo comprimento de x_A e que fornece os pesos para as métricas utilizadas, devolvendo assim um *ranking* personalizado. A função de *ranking* $h(\mathrm{x})$ é calculada por: $h(\mathrm{x}) = \mathrm{w}^T . \mathrm{x}_A$

Em **C-5** os autores também utilizam aprendizado de máquina, mais especificamente Linear Discriminant Analysis (LDA), para realizar o *ranking* das associações semânticas complexas. Essa técnica surgiu como evolução de J-2, pois não precisa mais que um usuário classifique manualmente um conjunto de treinamento para o aprendizado do algoritmo. Além disso, o algoritmo é capaz de evoluir com o tempo, aprendendo as mudanças de preferência do usuário a medida que o mesmo usa o sistema.

O algoritmo de C-5 funciona através de um ciclo contínuo de busca-*ranking*-*feedback*-refinamento. Após realizar uma busca, o algoritmo realiza o *ranking* dos resultados baseado no vetor de pesos w, o mesmo utilizado em J-2 e que é adquirido através dos gostos do usuário, e apresenta uma lista ordenada dos resultados. O usuário tem então a opção de informar se gostou da ordenação de um item na lista, através de um *feedback* positivo ou negativo em relação a ele. Dois vetores, um de positivos e um de negativos, são atualizados e utilizados no LDA para calcular o vetor de peso atualizado para aquele usuário, que será usado no *ranking* da sua próxima busca.

4.2.3 Estudos que usam abordagens probabilísticas

Em **C-7** é proposto um *ranking* probabilístico dos relacionamentos envolvendo uma dada entidade de entrada, usando a informação do grafo e também de um corpus do texto. Dada uma entidade de entrada e, para cada entidade alvo e_t e cada aresta r que conecte e e e_t, o peso da relação r entre e e e_t é calculado através da probabilidade $P(r, e_t|e)$,

$$P(r, e_t|e) \propto P(e_t) \times P(e|e_t) \times P(r|e, e_t), \quad (5)$$

cujos componentes são explicados abaixo.

$P(e_t)$ representa a probabilidade a priori da entidade alvo

e_t, e indica que, na ausência de qualquer outra informação, a entidade de entrada e tem chance maior de ter um relacionamento com uma entidade alvo popular do que com uma entidade rara. É estimada pela frequência relativa de relacionamentos em que e_t está envolvida.

P(e|e_t) representa a afinidade entre e e e_t, e indica que uma entidade alvo que tenha a maioria dos seus relacionamentos com a entidade de entrada é mais importante do que outra que tenha poucos relacionamentos ou relacionamentos muito fracos. É estimada pela frequência ponderada de relacionamentos entre e e e_t, em relação ao total de relacionamentos em que e_t está envolvida.

$P(r|e, e_t)$ representa a força do relacionamento r entre e e e_t, e indica que um relacionamento entre e e e_t que tenha maior suporte/evidência no *corpus* do texto é mais importante do que relacionamento com baixa evidência. É estimada pelo número de vezes que o relacionamento r entre e e e_t foi mencionado no texto dividido pelo número total de relacionamentos mencionados entre e e e_t.

Após calcular a Equação 5 para todos os relacionamentos, os resultados são apresentados em ordem decrescente.

4.2.4 Outros

Em **C-3**, os autores propõem um algoritmo capaz de sugerir ao usuário conceitos e relações mais importantes em uma dada ontologia, através de um algoritmo similar ao *PageRank*. No algoritmo proposto, eles definem quatro características para determinar a importância dos conceitos e relacionamentos: (i) um conceito é mais importante se existem mais relacionamentos conectados a ele; (ii) um conceito é mais importante se existem relacionamentos entre ele e outro conceito importante; (iii) um conceito se torna mais importante se existem relacionamentos de maior peso ligados a ele; e (iv) o peso de um relacionamento é maior se ele parte de um conceito importante. Uma vez que a importância do conceito e peso de um relacionamento dependem um do outro, o cálculo das importâncias das entidades e relacionamentos é realizado por um algoritmo iterativo.

Em **C-4**, os autores adaptam a métrica de *singularidade de uma propriedade* usada pelo algoritmo de *ranking* SemRank proposto em [3] e usam a técnica de *ativação por espalhamento*. A métrica do SemRank é estendida para medir a *singularidade de um recurso* relativa a todas as outras instâncias cuja classe pertence à mesma rede semântica. Essa métrica é usada junto com uma medida de relevância da entidade para inicializar o fator de ativação usado pela técnica de *ativação por espalhamento*. Nessa técnica, dadas as palavras-chave inseridas pelo usuário na consulta, o algoritmo faz a busca pelos nós no grafo através do cálculo de fator de ativação. Os autores definem um fator de decaimento para determinar quantos *espalhamentos* o algoritmo deve fazer, para assim atingirem uma quantidade razoável de entidades e relacionamentos que eles consideram relevantes à consulta. No fim, os resultados são ordenados em ordem de valor de ativação.

4.3 Características da abordagem: Q-2

A Tabela 3 contém uma organização dos estudos selecionados, de forma a responder as perguntas feitas em Q-2. No quesito da dependência de domínio, a grande maioria dos estudos, apesar de realizar seus experimentos em um único domínio, poderiam ser estendidos para outros domínios apenas mudando a base de dados utilizadas.

Quanto à dependência da *query*, há uma variação nos trabalhos, mas pode-se observar uma correlação entre a dependência da *query* e a personalização do *ranking*. Grande parte dos trabalhos que personalizam o *ranking* utilizam algum parâmetro, seja ele definido pelo usuário durante ou após a busca, ou usando o histórico de buscas do usuário para fazer a personalização, como em J-3. Isso permite que, em alguns casos, o conteúdo inserido na *query*, mesmo sendo o mesmo para diferentes usuários, implique em diferenças no resultado das funções de *ranking*.

A interface da *query* é, em grande parte, baseada na utilização do nome das entidades a fim de se obter as relações entre elas e construir o *ranking*. Dos trabalhos observados, apenas um, o C-5, apresenta a possibilidade de melhorar a função de *ranking* com o tempo. Esse fato se deve a que sua função de *ranking* refaz os cálculos dos pesos dos relacionamentos considerando a personalização do usuário.

Outra observação é que os algoritmos que fazem uma personalização do *ranking* e que utilizam técnicas de aprendizado de máquina, isto é, J-2 e C-5, não necessitam de uma configuração manual dos parâmetros de *ranking*. Ter uma configuração manual desses parâmetros por parte dos usuários é indesejável, pois requer que os usuários tenham um bom entendimento de toda a esquematização do sistema, o que é difícil para usuários inexperientes [8].

Há também, em alguns trabalhos, a preocupação com o contexto da *query*. Nesses casos, o contexto de uma busca tem influência no resultado de *ranking*, pois alguns relacionamentos entre duas entidades ganham mais significância dependendo da intenção (contexto) de interesse do usuário.

4.4 Características dos experimentos: Q-3

A Tabela 4 contém uma organização dos estudos selecionados, de forma a responder as perguntas feitas em Q-3.

Quanto aos experimentos, foi possível observar que existe uma variedade de bases disponíveis com uma grande variação de número de instâncias disponíveis em cada uma.

A qualidade do *ranking* foi uma preocupação recorrente nos trabalhos. Neles, procurou-se analisar se os melhores resultados gerados pelos algoritmos propostos eram melhores do que os outros estudos com quem eram comparados, ou se ao menos chegavam perto do desejo expressado pelos usuários. As métricas estatísticas usadas para avaliar os trabalhos, como a precisão e o ganho acumulado, servem para expressar matematicamente a relevância das técnicas propostas. Outra das métricas avaliadas foi a complexidade do tempo, que tem um peso significativo na experiência do usuário.

Todos os algoritmos de *ranking* dos estudos selecionados foram testados com usuários reais, sendo que alguns artigos não informaram a quantidade exata de humanos envolvidos nos experimentos. Esse tipo de testes demonstraram ser uma poderosa ferramenta de verificação do algoritmo.

Alguns trabalhos, como C-1, C-2, C-3 e C-4, não apresentaram dados ou apresentaram poucos dados sobre os experimentos. Uma possível explicação para isso é que, por se tratarem de trabalhos mais antigos, ainda estavam em fase de experimentação e proposição, não tendo ainda resultados concretos. Os primeiros artigos analisados não fizeram comparação com outros algoritmos de *ranking*. É possível que seja também devido ao fato de serem artigos mais antigos.

Table 3: Características da abordagem utilizada nos estudos selecionados nesta RS.

ID	Independente do domínio/domínio específico	Independente da *query*/Dependente da *query*	Interface da *query*	Melhora com o tempo	*ranking* personalizado	Config. manual parâmetros	Sensibilidade ao contexto
J-1	Independente	Dependente	Nome das entidades	Não	Sim	Sim	Sim
C-1	Independente, porém cada entidade deve ser manualmente associada a um contexto	Dependente	Nome das entidades, Contexto da *query*	Não	Sim	Sim	Sim
C-2	Independente	Dependente	Nome das entidades	Não	Sim	Sim	Não
C-3	Independente	Independente	Uma ontologia	Não	Não	Não	Não
C-4	Independente	Independente	Palavra-chave	Não	Não	Não	Não
J-2	Independente	Dependente	Nome das entidades	Não	Sim	Não	Não
C-5	Independente	Dependente	Nome das entidades	Sim	Sim	Não	Não
C-6	Independente	Dependente	Nome das entidades	Não	Sim	Sim	Sim
J-3	Independente	Dependente	Nome das entidades	Não	Sim	Sim	Sim
C-7	Independente	Independente	Nome da entidade	Não	Não	Não	Não

Table 4: Características dos experimentos realizados nos estudos selecionados nesta RS.

ID	Base de dados	Métricas	Comparação com	Nº de instâncias[8]	Nº de relações	Nº de humanos
J-1	SWETO[9]	Top k	Ninguém	800K	1,5M	5
C-1	Própria: domínio de artes	NI	Ninguém	NI	NI	NI
C-2	SWETO[9], US Senate Data	NI	Ninguém	NI	NI	NI
C-3	Schema*Web*[10]	Precisão dos melhores k, Top k, Coeficiente de correlação de Pearson	PageRank adaptado, AKTiveRank	NI	NI	5
C-4	Própria: domínio de eletrônica	NI	Ninguém	NI	NI	NI
J-2	Freebase linked-open-data[11]	Complexidade de tempo, Taxa de perda cumulativa, Qualidade do *ranking* & Ganho acumulado normalizado na posição k	SemDis[2], LtR_CA (SemDis + SVM)	185K	7K	20
C-5	Freebase linked-open-data[11]	Complexidade de tempo, Proporção de registros classificados entre os 10 melhores que foram rotulados como relevantes & Número de usuários que deram *feedback* na consulta k	SemDis[2], SVMLtR[9]	340K	590K	20
C-6	Próprio: domínio de música, finanças, terrorismo, esporte entre outros	Spearman footrule & Precisão dos melhores k	SemDis[2], SemRank [3], Lee's rank [18] Vidal's rank [23]	3K	70	5
J-3	Próprio: domínio de música, finanças, terrorismo, esporte entre outros	Spearman footrule & Precisão dos melhores k	SemDis[2], SemRank [3], Lee's rank [18] Vidal's rank [23]	3K	70	50
C-7	Wikipedia, KORE[10]	Comparação dos melhores 10	Baseline: relacionamentos mais populares	30M gerados, 21 avaliados	192M	2

5. DISCUSSÃO E CONCLUSÕES

Buscas que vasculham a *Web* Semântica, em especial considerando o relacionamento entre os dados, têm sido propostas, assim como soluções para melhor classificar os resultados a serem exibidos ao usuário. No entanto, ainda há espaço para melhorar a forma com que o motor de busca realiza a pesquisa e classifica os resultados, de maneira que estes sejam mais precisos, e tenham maior correlação com os interesses do usuário.

A grande maioria dos artigos que possuem personalização do resultado do *ranking* necessitam de uma configuração manual de parâmetros por parte dos usuários. De fato, os únicos dois métodos que não precisam de configuração manual são os que utilizam técnicas de aprendizado de máquina. Assim, o estudo da aplicação de técnicas de aprendizado de máquina se torna uma abordagem interessante para a área.

Somente um dos algoritmos presentes nessa RS apresentou a capacidade de aprender a preferência do usuário com o tempo, o que abre um espaço para pesquisas que desenvolvam algoritmos com essa característica.

Levando em consideração o uso de aprendizado de máquina e a evolução do algoritmo com o tempo, sugere-se o estudo de técnicas com essas duas características, como o aprendizado por reforço. Algoritmos de *ranking* que usam aprendizado por reforço têm sido aplicados com sucesso em motores de busca baseados em consulta a palavras-chave na *Web* tradicional, como em [22]. O algoritmo deles usa *feedback* do usuário para fazer o *ranking* de páginas *Web*. Como sugestão de trabalho futuro, seria interessante verifi-

[8]K equivale a mil, M a milhão e NI representa *Não Informado*

[9]http://knoesis.wright.edu/library/ontologies/sweto/

[10]http://www.schema*Web*.info

[11]http://schemaviz.freebaseapps.com/?domain=/fictional_universe

car a possibilidade de adaptação dessa técnica para a *Web* Semântica.

6. REFERENCES

[1] B. Aleman-Meza, I. B. Arpinar, M. V. Nural, and A. P. Sheth. Ranking documents semantically using ontological relationships. In *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*, pages 299–304. IEEE, 2010.

[2] B. Aleman-Meza, C. Halaschek-Wiener, A. Sheth, I. B. Arpinar, and C. Ramakrishnan. Ranking complex relationships on the semantic web. *IEEE Internet Computing*, pages 37–44, 2005.

[3] K. Anyanwu, A. Maduko, and A. Sheth. Semrank: Ranking complex relation search results on the semantic web. In *14th International Conference on World Wide Web*, pages 117–127, 2005.

[4] K. Anyanwu and A. Sheth. ρ-queries: Enabling querying for semantic associations on the semantic web. In *Proceedings of the 12th International Conference on World Wide Web*, pages 690–699, 2003.

[5] T. Bernes-Lee. Linked data - design issues. http://www.w3.org/DesignIssues/LinkedData.html, 2006. Acessado em 31-01-2017.

[6] S. Bhatia, A. Goel, E. Bowen, and A. Jain. Separating wheat from the chaff – a relationship ranking algorithm. *The Semantic Web: ESWC 2016 Satellite Events*, pages 79–83, 2016.

[7] P. Castells, M. Fernandez, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE transactions on knowledge and data engineering*, 19(2):261–272, 2007.

[8] N. Chen and V. Prasanna. Rankbox: An adaptive ranking system for mining complex semantic relationships using user feedback. In *Proceedings of the IEEE 13th International Conference on Information Reuse and Integration*, pages 77–84, 2012.

[9] N. Chen and V. K. Prasanna. Learning to rank complex semantic relationships. *International Journal on Semantic Web and Information Systems*, 8(4):1–19, Oct. 2012.

[10] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. Kore: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 545–554. ACM, 2012.

[11] A. Hogan, S. Decker, and A. Harth. Reconrank: A scalable ranking method for semantic web data with context. In *Proceedings of second international workshop on scalable semantic web knowledge base systems*, 2006.

[12] V. Jindal, S. Bawa, and S. Batra. A review of ranking approaches for semantic search on web. *Information Processing and Management*, 50(146):416–425, 2014.

[13] S. A. Kareem and P. M. Barnaghi. A context-aware ranking method for the complex relationships on the semantic web. In *International Conference on Advanced Language Processing and Web Information Technology*, pages 129–134. IEEE, 2007.

[14] B. A. Kitchenham. Systematic review in software engineering: Where we are and where we should be going. In *Proceedings of the 2Nd International Workshop on Evidential Assessment of Software Technologies*, pages 1–2. ACM, 2012.

[15] K. J. Kochut and M. Janik. Sparqler: Extended sparql for semantic association discovery. In E. Franconi, M. Kifer, and W. May, editors, *Proceeding of the 4th European Semantic Web Conference*, pages 145–159, 2007.

[16] F. Lamberti, A. Sanna, and C. Demartini. A relation-based page rank algorithm for semantic web search engines. *IEEE Transactions on Knowledge and Data Engineering*, 21(1):123–136, 2009.

[17] F. Lamberti, A. Sanna, and C. Demartini. A relation-based page rank algorithm for semantic web search engines. *IEEE Transactions on Knowledge and Data Engineering*, 21(1):123–136, 2009.

[18] M. Lee and W. Kim. Semantic association search and rank method based on spreading activation for the semantic web. In *2009 IEEE International Conference on Industrial Engineering and Engineering Management*, pages 1523–1527, 2009.

[19] X. Ning, H. Jin, and H. Wu. Rss: A framework enabling ranked search on the semantic web. *Information Processing & Management*, 44(2):893 – 909, 2008.

[20] A. J. Roa-Valverde and M.-A. Sicilia. A survey of approaches for ranking on the web of data. *Information Retrieval*, 17(4):295–325, 2014.

[21] A. K. Thushar. An RDF approach for discovering the relevant semantic associations in a social network. In *16th International Conference on Advanced Computing and Communications*, pages 214–220, 2008.

[22] V.Derhami, J.Paksima, and H. Khajeh. Web pages ranking algorithm based on reinforcement learning and user feedback. *Journal of AI and Data Mining*, 3(2):157–168, 2015.

[23] M.-e. Vidal, L. Rashid, L. Ibabez, J. Rivera, H. Rodrogiez, and E. Ruckhaus. A ranking-based approach to discover semantic association between linked data. In *The 2nd international workshop on inductive reasoning and machine learning for the semantic web*, volume 611, pages 18–29, 2010.

[24] V. Viswanathan and K. Ilango. Ranking semantic relationships between two entities using personalization in context specification. *Information Sciences*, 207:35–49, 2012.

[25] V. Viswanathan and I. Krishnamurthi. Ranking semantic associations between two entities – extended model. *Intelligent Information and Database Systems: 4th Asian Conference, ACIIDS 2012, Kaohsiung, Taiwan*, pages 152–162, 2012.

[26] W. Wei, P. Barnaghi, and A. Bargiela. Rational research model for ranking semantic entities. *Information Sciences*, 181(13):2823 – 2840, 2011.

[27] G. Wu, J. Li, L. Feng, and K. Wang. Identifying potentially important concepts and relations in an ontology. In *Proceedings of the 7th International Semantic Web Conference*, pages 33–49, 2008.

[28] S. Yumusak, E. Dogdu, and H. Kodaz. A short survey of linked data ranking. In *Proceedings of the 2014 ACM Southeast Regional Conference*, pages 48:1–48:4. ACM, 2014.