# Uma Proposta de *Entity Ranking* Baseada no Uso de Entidades como Objetos de Consulta

**Thiago C. Krug**[1], **Sergio L. S. Mergen**[1]

[1]Campus Alegrete - Universidade Federal do Pampa (UNIPAMPA)
97.546-550 – Alegrete – RS – Brasil

thiagockrug@gmail.com, sergiomergen@unipampa.edu.br

***Abstract.*** *This article explores the process of Entity Ranking based on relationships, with Wikipedia as a data source. We propose a graph model to represent the relationships between entities. The ranking is calculated based on the number of relationships between entities and their popularity. Furthermore, a graph compression technique is employed which is particularly useful for the type of structure used by Wikipedia. The effectiveness of the ranking is described through the analysis of answers for popular entities queries.*

***Resumo.*** *Este artigo explora o processo de Entity Ranking baseado em relacionamentos, tendo a Wikipedia como fonte de dados. É proposto um modelo de grafos para representar os relacionamentos entre as entidades. O ranking é calculado com base no número de relacionamentos entre as entidades e na popularidade que elas possuem. Além disso, é empregada uma técnica de compressão de grafos que é particularmente útil para o tipo de estrutura utilizada pela Wikipedia. A efetividade do ranking é descrita através de análises das respostas obtidas para consultas a entidades populares.*

## 1. Introdução

A área de *Entity Ranking* surgiu recentemente como um campo de pesquisa que visa recuperar entidades nomeadas como respostas a consultas. Exemplos de entidades incluem organizações, pessoas, localidades e datas. Esse problema difere da extração de entidades, onde o foco está na habilidade de identificar as entidades existentes em documentos usando técnicas de processamento de linguagem natural e treinamento a partir de grandes coleções de documentos. Já o processo de *Entity Ranking* foca na recuperação de uma lista ordenada de entidades que possuam relação a uma consulta por palavras-chave.

O interesse por essa área de pesquisa se acentuou a partir da constatação de que uma considerável fração das buscas na *Web* referenciavam entidades nomeadas [Paşca 2007]. Além disso, existem domínios de sites na *Internet* que podem ser considerados grandes repositórios de entidades, o que encoraja a criação de aplicações que utilizem esse tipo de informação. Um dos exemplos mais significantes é a *Wikipedia*.

A forma como os dados estão estruturados na *Wikipedia* a torna um recurso poderoso que pode ser explorado para muitos tipos de análise de dados, como desambiguação de palavras [Mihalcea 2007], recuperação de informação e resposta a consultas [Ahn et al. 2004]. Alguns dos trabalhos de pesquisa que utilizam a *Wikipedia* como fonte de dados trata do problema conhecido como *Entity Ranking*, que envolve encontrar as melhores entidades como resposta a consultas. Ao contrário das técnicas tradicionais

de recuperação de informação na *Web*, onde as respostas são páginas HTML publicadas na *Web*, soluções baseadas em *Entity Ranking* trazem como respostas as entidades publicadas na *Wikipedia*.

Neste artigo, será abordado um problema relacionado ao de *Entity Ranking*, onde o objetivo envolve descobrir entidades que sejam relevantes a uma entidade de consulta, e não a um conjunto de palavras chave. Essa variação do problema original remete a situações em que o usuário já conhece um tópico, e deseja conhecer outros tópicos que estejam relacionados.

Para lidar com essa questão, o artigo propõe uma abordagem que explora a existência de relacionamentos entre as entidades para o cálculo da relevância. O cálculo leva em consideração a intimidade entre entidades relacionadas e a popularidade delas perante todas as demais. Além disso, é apresentado um mecanismo de extração de relacionamentos baseado na *Wikipedia*, onde a estrutura de categorias é usada para identificar os relacionamentos entre as entidades. O artigo também relata como os relacionamentos podem ser descritos em um grafo de entidades, e apresenta uma técnica de compressão que pode ser usada para reduzir o custo de espaço para representação do grafo.

Este artigo está organizado da seguinte forma: A Seção 2 apresenta a abordagem de *ranking* proposta, que compreende a noção de intimidade e popularidade das entidades. A Seção 3 descreve o mecanismo usado para extrair tanto entidades como relacionamentos a partir da *Wikipedia*. Na Seção 4 são descritos experimentos realizados sobre a coleção INEX, que contempla todas as páginas da *Wikipedia* americana de 2007. Os experimentos demonstram como a abordagem proposta ordena os resultados para algumas entidades de consulta, e apresentam a taxa de compressão que foi obtida sobre o grafo de entidades. Os trabalhos relacionados são descritos na Seção 5.

## 2. *Entity Ranking* Baseado em Relacionamentos

O cálculo do *ranking* proposto neste artigo considera a existência de relacionamentos entre as entidades. A partir da análise desses relacionamentos, pretende-se determinar quais dessas entidades são mais relevantes, tanto em relação a entidade especificada na consulta quanto em relação ao conjunto total de entidades mapeadas.

Os relacionamentos entre as entidades podem ser representados por uma estrutura de grafo, onde as entidades são vértices e os relacionamentos entre as entidades são as arestas, conforme a Definição 1.

**Definição 1** *(Grafo de Entidades) Considere que $G = \{V, A, T\}$ seja um grafo de entidades orientado e valorado, onde o conjunto de vértices $V$ é composto pelas entidades $e$, enquanto o conjunto de arestas $A$ é composto por relacionamentos entre essas entidades. Ainda, considere que $r = (e_i, e_j, tipo)|r \in A$ seja um relacionamento, onde $e_i \in V$ é uma entidade de origem do relacionamento, $e_j \in V$ é uma entidade de destino do relacionamento, e $tipo \in T$ caracteriza o tipo de relacionamento.*

A Figura 1 ilustra um grafo que satisfaz essa restrição. O universo de dados capturado no modelo contempla oito entidades, cinco tipos de relacionamento e dez relacionamentos. Como as arestas não são orientadas, o modelo é capaz de representar apenas relacionamentos simétricos. Outros tipos de relacionamentos, apesar de possíveis, não são considerados nesse trabalho.

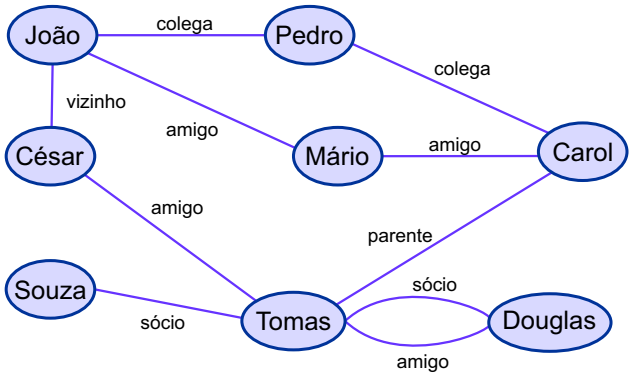


**Figura 1. Grafo de Exemplo Capturando Relacionamentos entre Oito Entidades**

Dada uma consulta por entidade, o retorno será a lista de entidades relevantes ao objeto de consulta. Neste trabalho, essa lista é determinada conforme a Definição 2.

**Definição 2** *(Lista de Resposta) Considere que* $G = \{V, A, T\}$ *seja um grafo de entidades orientado e valorado, e* $e_c \in V$ *seja um objeto de consulta. Nesse caso, a lista de resposta será composta por entidades* $e_r \in V$*, onde* $\exists r \in A | (r.e_o = e_c \wedge r.e_d = e_r) \vee (r.e_o = e_r \wedge r.e_d = e_o)$.

A definição acima parte da intuição de que uma entidade é relevante a uma outra entidade se existirem relacionamentos entre elas. Por exemplo, na Figura 1, dada uma consulta pela entidade ”Tomas”, a lista de resposta seria composta por quatro entidades: ”César”, ”Souza”, ”Carol”e ”Douglas”.

O próximo passo envolve determinar a ordem de exibição das entidades de resposta, de modo que as mais relevantes apareçam primeiro. Neste artigo, considera-se que as entidades mais relevantes a um objeto de consulta sejam aquelas que compartilhem mais relacionamentos com o objeto de consulta. Essa regra procura medir o nível de afinidade entre duas entidades. Quanto mais relacionamentos em comum, maior é a afinidade.

De acordo com esse critério, ”Douglas”seria a entidade mais próxima a ”Tomas”, uma vez que elas compartilham mais relacionamentos (de amizade e sociedade). As demais possibilidade de resposta teriam menos afinidade com ”Tomas”, já que compartilham com o objeto de consulta apenas um relacionamento cada (”César”/amigo, ”Souza”/sociedade e ”Carol”/família).

Caso a consulta fosse sobre ”João”, haveria três entidades de resposta. No entanto, o nível de afinidade com o objeto de consulta é o mesmo. Nesse caso, o critério de desempate passa a ser a popularidade das entidades. O cálculo de popularidade adotado neste artigo é discutido na próxima seção.

### 2.1. Cálculo de Popularidade

O exemplo clássico de *ranking* por popularidade é conhecido como *PageRank*, onde a relevância de páginas *Web* é usada para ordenar os resultados de buscas por palavra chaves [Brin and Page 1998]. O algoritmo do *PageRank* trata a *Web* como um grafo, onde as páginas são os vértices e os *links* entre páginas são arestas orientadas. O *ranking* é determinado pela popularidade das páginas, que é calculada de acordo com o número de

*links* que apontam para elas. Além disso, a popularidade de uma página é proporcional a popularidade das páginas que apontam para ela. O cálculo é baseado na probabilidade de que uma página seja acessada aleatoriamente, dado o conjunto total de páginas e seus respectivos *links*. O cálculo aproximado da probabilidade é realizado através de simulações de navegação, onde um usuário fictício percorre as páginas existentes, seguindo os *links* internos aleatoriamente.

A intuição da popularidade é válida quando aplicada ao *ranking* de páginas, mas perde um pouco do sentido se analisada no contexto das entidades conforme modelado neste artigo. Afinal, no grafo *Web* de páginas, os *links* são orientados, e representam ”indicações”de que uma página possa ser útil. Naturalmente, quanto mais indicações, mais útil a página deve ser. Já no modelo de entidades proposto, os relacionamentos apresentam associações simétricas, onde a importância das entidades participantes não pode ser representada.

Porém, mesmo partindo de um modelo de grafo semanticamente diferente, um conceito geral de popularidade pode ser explorado para atribuição da importância das entidades. Sendo assim, considera-se que a popularidade seja relativa ao número de arestas das entidades. Ou seja, entidades que possuam mais relacionamentos são mais populares. Além disso, a popularidade de uma entidade pode ser propagada para as entidades relacionadas. Ou seja, entidades que se relacionam com entidades populares serão proporcionalmente mais populares.

Para ilustrar, considere novamente a Figura 1. Com base nos relacionamentos indicados, ”Tomas”seria a entidade mais popular, dado o número de arestas conectadas a ela. As entidades ”César”, ”Mário”e ”Pedro”aparecerem empatadas, com dois relacionamentos cada. No entanto, ”César”seria considerada mais popular, visto que ela é a única das três que se relaciona com a entidade mais popular.

Com base nessas considerações, propõe-se um algoritmo de propagação de popularidade baseado em um número fixo de iterações. Em cada iteração, a popularidade de uma entidade é calculada como a soma de sua própria popularidade e a popularidade das entidades com quem ela se relaciona. Essa soma leva em consideração as popularidades atribuídas na iteração anterior. É importante destacar que antes da primeira iteração é atribuída uma popularidade igual a todas entidades, para que nenhuma seja favorecida artificialmente. Quanto maior o número de iterações usado, mais distante a popularidade de uma entidade será propagada. Além disso, a popularidade de uma entidade é propagada a um vizinho um número de vezes igual ao número de relacionamentos conectando as entidades. Essa medida visa reforçar o fato de que a propagação deva ser mais acentuada se duas entidades estão mais fortemente relacionadas.

## 3. Extração de Entidades e Relacionamentos

Nesse artigo adota-se uma abordagem simples, tanto para identificação das entidades quanto dos relacionamentos, explorando a estrutura de categorias da *Wikipedia*. De acordo com o *template* de formatação da *Wikipedia*, o rodapé dos artigos possui uma listagem das categorias que estão relacionadas ao tema do artigo em questão. Ao clicar em uma categoria, o usuário é redirecionado a uma página de categoria, que exibe todos os artigos que pertençam a essa categoria.

Como cada artigo da *Wikipedia* versa a respeito de um assunto em particular, pode-

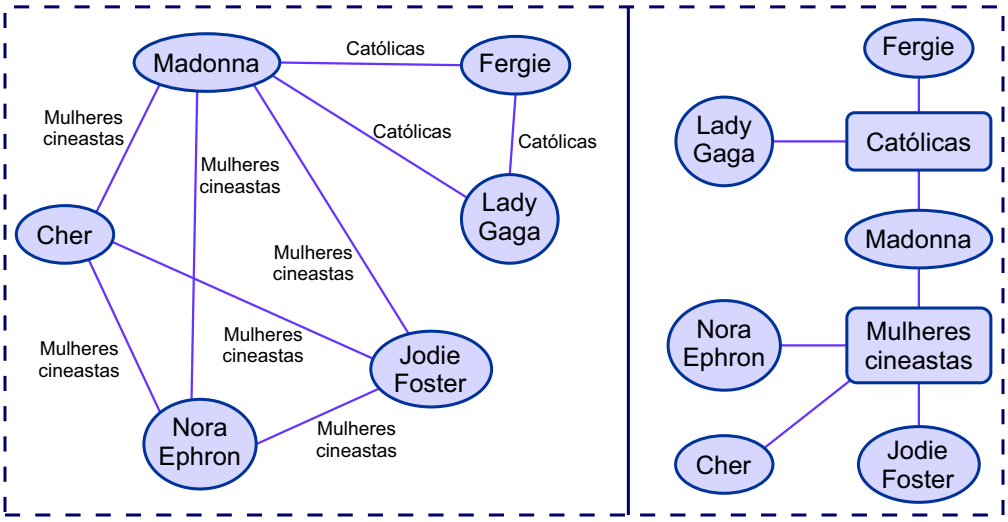


**Figura 2. Grafo de Relacionamentos Modelando o Universo de Dados de Exemplo**

se concluir que cada artigo é uma entidade por si só. Já em uma página de categorias, é possível considerar que os artigos (entidades) ali mencionados possuam relacionamentos entre si, onde o nome do relacionamento vem a ser o nome da categoria. A Figura 2 à esquerda apresenta um grafo gerado a partir da *Wikipedia* de acordo com a abordagem descrita. Para fins de ilustração, apenas uma parte do universo de dados está sendo representada.

Um ponto que vale a pena destacar a partir desse grafo é a presença de cliques, ou seja, um subconjunto de vértices onde cada dois vértices do subconjunto são conectados por uma aresta. Na verdade, verifica-se de um tipo de clique mais específico, onde as arestas possuem o mesmo rótulo. Cada clique neste formato corresponde aos relacionamentos extraídos a partir de uma página de categoria.

Neste artigo, exploramos a existência desse tipo de estrutura para aplicar uma técnica que visa comprimir o grafo. Em suma, a técnica proposta tem por objetivo eliminar os cliques existentes. Dado um clique, o primeiro passo da técnica envolve criar um vértice especial, chamado de vértice de tipo. Em seguida, as arestas que pertenciam ao clique são removidas. Por último, são criadas arestas conectando os vértices que participavam do clique ao novo vértice de tipo.

Pode-se ver que essa técnica aumenta o número de vértices do grafo e reduz o número de arestas. Para compreender esse comportamento, considere o exemplo simples em que o universo de dados compreende $n$ entidades que partilham uma categoria. Para representar esse fato, o grafo original demandaria $n$ vértices e um número de arestas equivalente ao intervalo $n * (n-1)/2$. Já o grafo modificado demandaria $n+1$ vértices (um vértice extra para representar o tipo) e $n$ arestas. A Figura 2 à direita serve como constatação dessa afirmação. Nesse exemplo, o grafo adaptado possui dois vértices a mais e duas arestas a menos do que o grafo original.

Em ambos exemplos citados acima, a suposição é que existam poucas categorias compartilhadas pelas entidades. Em um cenário mais realista, como no domínio de sites da *Wikipedia*, existem diversas ilhas de entidades que partilham categorias em comum.

Além do mais, o número de entidades que compartilham uma mesma categoria é maior. A união desses dois fatores faz a redução do número de arestas compensar o aumento do número de vértices. A Seção 4 apresenta experimentos que demonstram a economia de espaço que essa técnica de compressão pode representar.

## 4. Análise Experimental

Os experimentos relatados neste artigo tem propósito duplo. Em primeiro lugar, será feita uma análise do custo em espaço comparando o grafo de relacionamentos puro com o grafo de relacionamentos comprimido. Além disso, serão analisados os resultados da ordenação usando o algoritmo de *Entity Ranking* proposto.

Para ambas as análises é utilizada a coleção INEX 2007[1], que contem todas as páginas da *Wikipedia* americana no ano de 2007. A coleção é usada como *benchmark* para a realização de experimentos de *Entity Extraction*. Ao todo, 659388 verbetes e 115625 categorias são indexados. Além disso, são disponibilizadas consultas prontas juntamente com a resposta esperada. Como as consultas são compostas por lista de palavras chave e categorias, não se pode aproveitá-las nos experimentos.

Os resultados atingidos são apresentados a seguir.

### 4.1. Análise do Consumo de Memória

Todos verbetes, categorias e relacionamentos entre verbetes foram extraídos a partir do INEX e alimentados em uma base de dados MySQL. Para fins de comparação, duas estruturas de banco foram utilizadas, sendo que uma modela o grafo de relacionamentos puro enquanto a outra modela o grafo comprimido.

O modelo não comprimido ocupa cerca de 10,13 GB de espaço físico, enquanto o modelo comprimido ocupa cerca de 175,48 MB, o que caracteriza uma melhora de 83,08% na economia de espaço em disco. A justificativa para tamanha diferença pode ser compreendida ao analisar-se os dados da coleção. Mais de 26785 categorias são associadas com ao menos 10 verbetes. Além disso, algumas categorias estão associadas a um número de verbetes bastante elevado. Por exemplo, uma categoria chega a possuir associações com 4534 verbetes. Quanto maiores forem os cliques no grafo, maior o ganho em se optar pela versão compacta. Curiosamente, 40024 categorias não estão associadas a nenhum verbete.

### 4.2. Análise do Algoritmo de Ordenamento

Com o intuito de verificar o funcionamento do algoritmo de ordenamento de entidades, foram criados cinco objetos de consulta, compostos pelas seguintes entidades: ”Superman”, ”IBM”, ”Madonna”, ”Friends”e ”Counter-Strike”.

Para cada objeto de consulta, foram recuperadas as entidades de resposta nas posições 1, 2, 3, 4, 5, 31, 51, 81, 131, 211, nesta mesma ordem. Os cinco primeiros resultados caracterizam entidades bastante relacionadas ao objeto de consulta, sendo que o grau de relacionamento a princípio não difere muito entre eles. Os cinco últimos resultados caracterizam entidades pouco relacionadas ao objeto de consulta, sendo que o grau de relacionamento a princípio cai bastante conforme se usa entidades de ordem inferior.

[1]http://www-connex.lip6.fr/ denoyer/wikipediaXML/

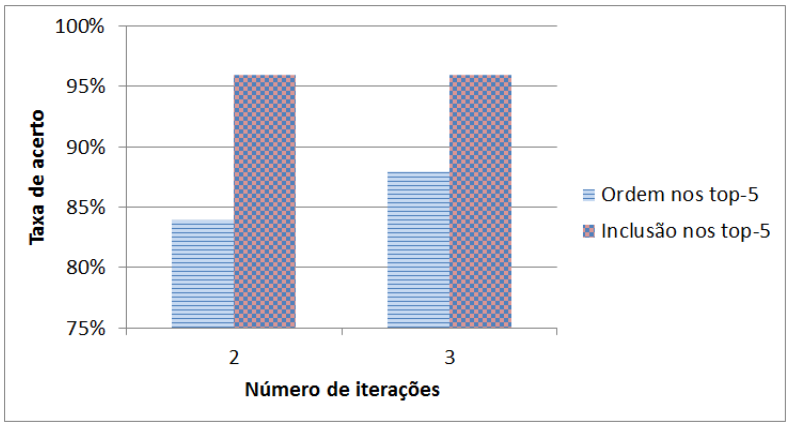


**Figura 3. Resultados da Propagação de Popularidade**

Um dos objetivos do experimento é medir se usuários concordam com a forma com que as entidades relacionadas foram ordenadas. Para realizar essa medição, doze usuários responderam a um questionário online[2]. Para cada objeto de consulta, o questionário pedia ao usuário que ordenasse as dez entidades relacionadas de acordo com a relevância. Para que a avaliação não fosse tendenciosa, as entidades foram dispostas em ordem alfabética.

Foram criados dois critérios de avaliação dos resultados, cuja definição e resultados obtidos são apresentados a seguir:

**Inclusão nos top-5** Verifica quantas das cinco entidades melhor ordenadas pelos usuários estão contidas nos top-5. Após realizar a média das respostas de todos os usuários, chegou-se a uma taxa de 63%. Esse resultado é um indicativo de que a estratégia de ordenação adotada é capaz de separar entidades bastante relacionadas das pouco relacionadas, para a maioria dos casos.

**Ordem nos top-5** Verifica quantas das cinco entidades melhor ordenadas pelos usuários estão na mesma ordem no top-5. Após realizar a média das respostas de todos os usuários, chegou-se a uma taxa de 17%. Esse resultado pouco expressivo pode ser interpretado pela dificuldade natural de determinar a relevância em alguns casos. Por exemplo, para o objeto de consulta ”Superman”, ”Flash”e ”Batman”são possibilidade de resposta. No entanto, os usuários tiveram dificuldade de determinar qual dessas possibilidades deveria ser melhor ordenada.

Outro experimento realizado analisou o desempenho do algoritmo de ordenação quando a popularidade é propagada a mais níveis de iteração. A Figura 3 apresenta os resultados obtidos, comparando a ordem alcançada usando um nível de propagação com a ordem alcançada usando dois e três níveis de propagação.

Conforme apresentado na Figura 3, a ordem e a inclusão não mudam drasticamente em comparação com a ordem alcançada com um nível de propagação. Isso ocorre principalmente porque o método de ordenação leva em consideração a afindade em primeiro lugar, e em segundo a popularidade. Ou seja, se o número de relacionamentos for igual, a ordem dos resultados não muda quando se alterar o número de iterações de

[2]http://fluidsurveys.com/surveys/thiago-krug/nivel-de-relacionamento-entre-entidades/

propagação de popularidade.

Também é possível constatar que existem poucas diferenças usando dois ou três níveis de propagação, o que também pode ser explicado pela forma como o escore de ordenamento é calculado. No entanto, convém destacar que algumas mudanças na ordem interessantes ocorreram quando passou a se usar mais níveis de propagação. É citado como exemplo o objeto de consulta ”Superman”. Usando um nível de propagação, a entidade relacionada ”Kon-El”aparece na 14ª posição. Seria desejável que essa entidade obtivesse uma relevância maior, visto que se trata de um personagem que pertence a dinastia do ”Superman”, o que ocorre quando se usa mais níveis de propagação.

## 5. Trabalhos Relacionados

O problema de *Entity Ranking* se assemelha bastante com a proposta deste artigo. Enquanto o *Entity Ranking* tradicional trata de encontrar entidades relacionadas a uma consulta por palavras-chave (e um conjunto opcional de entidades e categorias), a proposta de trabalho do artigo já parte de uma entidade existente, e pretende descobrir as entidades relevantes. Em seguida são discutidos alguns dos trabalhos de *Entity Ranking* existentes.

O trabalho de [Zhu et al. 2007] recupera primeiro todas as entidades que aparecem em páginas onde as palavras chave da consulta também aparecem. A relevância da entidade é computada com base na proximidade entre essa entidade e as palavras chave. Em seguida, as entidades recuperadas são filtradas caso nenhuma de suas categorias coincida com as categorias usadas na consulta.

Em [Kaptein and Kamps 2013], a busca é realizada através de uma série de scores, sendo que um deles compara as categorias da consulta e as categorias pertencentes a cada entidade indexada. A comparação é baseada na distância entre as categorias, calculada através da métrica KL-Divergence. Para o cálculo da divergência, são considerados os termos de todas as entidades que pertencem às categorias comparadas. A relevância de uma entidade é computada em função da menor distância entre qualquer das categorias pertencentes a pesquisa e qualquer das categorias pertencentes a entidade.

Já em [Vercoustre et al. 2008], a distância entre os dois conjuntos de categorias é calculada através de uma função $\frac{cat(Q) \cap cat(E)}{cat(E)}$ que mede a razão das categorias da consulta e da entidade que intersectam com relação ao total de categorias da entidade. Essa função compartilha da intuição que é usada neste artigo, a de que o compartilhamento de categorias indica relevância entre entidades. No entanto, o artigo dá mais importância ao número de compartilhamentos do que a proporção relativa deles. Além disso, tanto o trabalho de [Vercoustre et al. 2008] como os demais citados nesta seção possuem métricas adicionais, que utilizam as palavras chave da consulta para o cálculo, enquanto a proposta do artigo é completamente baseada no conceito de compartilhamento de categorias.

Como o foco é voltado ao uso de categorias, a possibilidade de compactar grafos que possuam essa característica se torna relevante. Dentro desse contexto, grafos baseados na *Web* possuem características próprias que podem melhorar as taxas de compressão. Em geral, esse tipo de grafos representa relações entre páginas HTML derivadas a partir de *hyperlinks*. Com base em estudos, descobriu-se que a criação desses relacionamentos obedece a alguns padrões. Por exemplo, páginas dentro de um domínio costumam citar a si próprias. Isso permite com que a as páginas sejam numeradas de acordo com a ordem

lexicográfica de URL, para que páginas de um mesmo domínio possuam identificadores próximos [Bharat et al. 1998, Blandford et al. 2003].

Outro exemplo considera que muitas páginas incluem *links* para o mesmo conjunto de páginas. A partir dessa observação, [Kumar et al. 1999] propôs uma variação de grafos bipartidos para representar o conjunto de páginas que citam e o conjunto de páginas citadas. Ocorrências desses grafos alimentam uma base de conhecimento, que pode ser usada para acelerar buscas e para processos de mineração de dados.

Seguindo a mesma linha, [Adler and Mitzenmacher 2001] propôs técnicas de similaridade que encontram páginas com listas de adjacência semelhantes. Após descobertas, a lista de adjacência de uma página é substituída por uma referência à página similar, juntamente com operações de edição para representar os pontos em que as duas listas originais divergiam.

Em [Buehrer and Chellapilla 2008], os conjuntos de páginas largamente citadas geram vértices estrela. As páginas que citam esse conjunto de páginas passam a citar esse vértice agrupador, o que reduz o número de arestas do grafo. Esse tipo de compressão é semelhante ao que nós empregamos nesse artigo, no sentido de que vértices especiais são gerados.

Uma distinção importante entre os trabalhos de compressão citados e o apresentado neste artigo é o fato de que as propriedades que possibilitam uma compressão precisam ser descobertas, como os grafos bipartidos ou listas de adjacência semelhantes. Já no contexto explorado no artigo, essas propriedades podem ser automaticamente derivadas durante a extração dos relacionamentos.

## 6. Conclusões

Este artigo explora uma forma diferente de *Entity Ranking*, em que o objeto de consulta é uma entidade, e o objetivo é encontrar a lista de entidades que estão relacionadas à entidade pesquisada. Os resultados obtidos mostram que os conceitos de afinidade e popularidade das entidades são parâmetros úteis na ordenação. O primeiro deles consegue realizar a divisão entre entidades bastante próximas ao objeto de consulta e entidades que são menos próximas. O segundo mostrou-se particularmente interessante em alguns casos específicos, especialmente quando a popularidade era propagada por mais de dois níveis no grafo de relacionamentos. Nesses casos, entidades reconhecidamente mais conhecidas ultrapassaram entidades menos conhecidas.

No entanto, o impacto da popularidade no resultado é reduzido, uma vez que o principal critério de ordenação é o nível de afinidade. Dessa forma, pretende-se estudar outras formas de usar esses dois valores para chegar a um escore final. Além disso, outro problema a ser estudado envolve a extração de relacionamentos entre entidades a partir de texto em linguagem natural. A presença de dados extraídos de outros tipos de fontes de dados pode inclusive levar a criação de um novo critério de ordenação, baseado na importância inferida ao método de extração que gerou o relacionamento. Essa regra partiria da intuição de que duas entidades que apareçam dentro de uma mesma sentença tenham um grau de afinidade diferente do que entidades que partilhem de uma mesma categoria na *Wikipedia*.

Para finalizar, é enfatizado o fato de que a estrutura dos relacionamentos na *Wi-*

*kipedia* permitiu que se alcançasse altos graus de compressão dos dados sem perda de informação. O próximo passo é analisar se a técnica de compressão usada segue sendo útil quando novos critérios de ordenação forem usados. Caso a ordenação empregue características específicas no relacionamento entre entidades (como o nível de afinidade), a técnica de compressão precisaria ser modificada para evitar que essas características específicas sejam perdidas.

## Referências


Adler, M. and Mitzenmacher, M. (2001). Towards compressing web graphs. In *DCC*, pages 203–212. IEEE Computer Society.

Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., and Schlobach., S. (2004). Using wikipedia at the trec qa track. In *Proceedings of TREC 2004*.

Bharat, K., Broder, A., Henzinger, M. R., Kumar, P., and Venkatasubramanian, S. (1998). The connectivity server: Fast access to linkage information on the Web. In *Proceedings of the 7th International World Wide Web Conference (WWW-7)*, pages 469–477, Brisbane, Australia.

Blandford, Blelloch, and Kash (2003). Compact representations of separable graphs. In *SODA: ACM-SIAM Symposium on Discrete Algorithms (A Conference on Theoretical and Experimental Analysis of Discrete Algorithms)*.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V.

Buehrer, G. and Chellapilla, K. (2008). A scalable pattern mining approach to web graph compression with communities. In Najork, M., Broder, A. Z., and Chakrabarti, S., editors, *WSDM*, pages 95–106. ACM.

Kaptein, R. and Kamps, J. (2013). Exploiting the category structure of wikipedia for entity ranking. *Artif. Intell*, 194:111–129.

Kumar, S. R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Extracting large-scale knowledge bases from the web. In *Proceedings of the 25th VLDB Conference*.

Mihalcea, R. (2007). Using wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL HLT*, volume 2007, pages 196–203.

Paşca, M. (2007). Weakly-supervised discovery of named entities using web search queries. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 683–690, New York, NY, USA. ACM.

Vercoustre, A.-M., Pehcevski, J., and Thom, J. A. (2008). Focused access to xml documents. chapter Using Wikipedia Categories and Links in Entity Ranking, pages 321–335. Springer-Verlag, Berlin, Heidelberg.

Zhu, J., Song, D., and Rüger, S. M. (2007). Integrating document features for entity ranking. In Fuhr, N., Kamps, J., Lalmas, M., and Trotman, A., editors, *INEX*, volume 4862 of *Lecture Notes in Computer Science*, pages 336–347. Springer.