TRANFORMAÇÃO DE TEXTOS EM GRAFOS CONCEITUAIS

Aluno: Edson Leon Serva Junior
Orientador: Madiagne Diallo

Introdução

A globalização permitiu uma maior facilidade de comunicação, possibilitando a instantânea troca de informações entre o mundo inteiro. O advento da internet possibilitou um acervo considerável de documentos eletrônicos e consequentemente uma grande variedade de informação que pode ser obtida através desse meio. Contudo, a busca por algo específico se tornou difícil diante do desafio de saber selecionar, no meio de tantos documentos, quais seriam relevantes para o usuário. Com isso, torna-se evidente a necessidade de criação de mecanismos de busca mais eficientes e precisos, ou seja, suficientemente rápidos e que retornem documentos pertinentes a query utilizada pelo usuário.

Atualmente muitos mecanismos de busca utilizam palavras-chave para representação de documentos. Um exemplo de mecanismo de busca que utiliza esse método é o Lucene. Porém, na área de recuperação de informações (RI), por exemplo, têm sido utilizado os grafos conceituais (GC) como alternativa às palavras-chave, o que otimiza a precisão do resultado, uma vez que são utilizados mais tipos de elementos textuais na representação.

Alguns modelos para a obtenção da relevância de documentos podem ser encontrados na literatura destacando-se coeficiente de Dice (CD), coeficiente de Jaccard e coeficiente de Cosine.

A equação abaixo mostra um modelo matemático (CD) para a obtenção da relevância de um documento a partir de uma query:

$$S_{D_1,D_2} = \frac{2n(D_1 \cap D_2)}{n(D_1)+n(D_2)}$$ Onde $n(D_i)$ é o número de termos em D_i e $n(D_i \cap D_j)$ é o número de termos que ambos os documentos possuem em comum.

Devido à sua simplicidade e normalização, o CD será a base do método de comparação entre dois textos representados como GC, sendo acrescentados alguns elementos devido a natureza bipartida dos GC (conceitos e relações) .

Objetivos

Estudar o conceito de GC, a representação do conteúdo de um texto como GC e medir o grau de similaridade entre dois textos representados como GC através do método proposto por [1]. Pretende-se também mostrar algumas aplicabilidades desse método, como por exemplo, em um sistema de RI. Após a fase de estudo, tem-se por objetivo programar tal modelo, testá-lo e validá-lo.

Metodologia

Neste método, as relações consideradas nos GC são de tipos básicos, como atributo, sujeito, objeto, etc. Ilustrando:

A frase *John loves Mary* é representada pelo grafo [John] ← (subj) ← [love] → (obj) → [Mary] e não por [John] ← (love) → [Mary]. [1]

Na construção de uma representação de uma frase em GC, este método utiliza um "part- of- speech tagger" (classificador de partes do discurso), um analisador sintático e um

analisador semântico. Primeiramente, o "part- of- speech tagger" fornece cada palavra com um papel sintático. Depois o analisador sintático gera sua representação estruturada, e por último, o analisador semântico gera um ou mais GC por cada estrutura sintática.

Para comparar 2 textos ou 2 frases, por exemplo um documento e a query do usuário, primeiramente suas representações como GC são construídas. Cada frase ou texto pode ser representado por um conjunto de GC, como por exemplo, em frases longas ou textos com muitas frases. Na área de RI, uma das principais medidas de similaridade entre o grafo da query utilizada pelo usuário e os grafos do documento é se o grafo da query utilizada pelo usuário está completamente contido no grafo do documento. Se estiver, o resultado obtido é relevante para a query usada. Em outras palavras, esse critério significa que o conteúdo de um documento deve ser mais específico do que a consulta para que o resultado seja satisfatório.

O algoritmo de comparação é constituído por 2 partes principais:

1- Encontrar a interseção entre os 2 conjuntos de GC.
2- Medir a similaridade entre os 2 conjuntos de GC como o tamanho relativo de cada um de seus grafos de interseção.

O grafo de interseção G_C apresenta todos os nós de conceito que aparecem nos GC originais G_1 e G_2 e todos os nós de relação que aparecem tanto em G_1 quanto em G_2 e relacionam os mesmos nós de conceito.

A similaridade entre 2 GC é uma combinação de 2 valores: a similaridade conceitual S_C e a similaridade relacional S_C . Abaixo as fórmulas: [1]

$$s_c = \frac{2n(G_c)}{n(G_1)+n(G_2)}$$ Onde n(G) é o nº de nós de conceito de um Grafo G. S_C varia de 0 a 1

$$s_r = \frac{2m(G_c)}{m_{G_c}(G_1)+m_{G_c}(G_2)}$$ onde $m(G_c)$ é o nº de arcos do grafo G_C e $m_{G_c}(G_1)$ é o nº de arcos do vizinho imediato do grafo GC no grafo GI

$s = s_c \times (a + b \times s_r)$. Onde *a* indica o valor da fração que a similaridade geral representa da similaridade conceitual.

$$a = \frac{2n(G_c)}{2n(G_c)+m_{G_c}(G_1)-mG_c(G_2)}$$ Onde n(Gc) é o nº de nós de conceito em GC e $m_{G_c}(G_1)+m_{G_c}(G_2)$ é o nº de nós de relação em G1 e G2 que são conectados aos nós de conceito aparecendo em Gc. Quando $S_r = 1$, $b = 1 - a$.

Conclusões

Na área de RI, a representação de textos como GC são uma alternativa mais eficaz e inteligente do que a representação por palavras-chave. Isso é refletido na melhora da precisão do processo de recuperação através de uma melhor classificação dos resultados. O projeto encontra-se ainda em andamento, em fase de estudo de como programar o modelo proposto por [1], a fim de testá-lo e depois validá-lo.

Referências

1 - Gomez, Manuel Montes y; López, Aurélio; Gelbukh, Alexander F**. Information Retrieval with Conceptual Graph Matching.** 11th International Conference on Database and Expert Systems Applications, Londres, v. 1873, p. 312–321, set.2000.