Propuesta de Trabajo de Tesis
Magister en Ingeniería Informática - Agosto 2014

Information Retrieval and Web Search: Exploring Low Cost Methods for Retrieving Semantically Similar Terms and Documents

El Contexto del Problema

Durante las últimas décadas, el amplio uso de la Web está basado en métodos de recuperación de información desde una colección grande de documentos de texto (más de 8 billones de páginas web indexadas por Google). Esta recuperación debe ser rápida y efectiva, es decir, los documentos recuperados deben ser relevantes a la query del usuario y a la vez éstos deben ser recuperados de una manera rápida.

El indexamiento semántico latente (LSI, latent semantic indexing) es uno de los métodos que considera la estructura semántica latente en los datos, por lo que ofrece ventajas sobre los modelos de recuperación basados en palabras claves o matching de términos. LSI está basado en una técnica estadística denominada Descomposición en Valores Singulares (SVD, singular value decomposition).

Aunque LSI genera mejores resultados que los métodos tradicionales basados en palabras claves para recuperar información relevante desde la Web, sus principales desventajas radican en: a) la complejidad en tiempo del método LSI, el cual es $O(m^2n)$, donde m es el número de términos y n es el número de documentos; b) el espacio de conceptos no es interpretable directamente; c) el número óptimo de dimensiones a retener usualmente se identifica por prueba y error, lo cual es muy consumidor de tiempo.

Objetivo de la Tesis

El objetivo principal de este trabajo de tesis es experimentar con métodos de machine learning (ML) adaptados que puedan aproximar los resultados obtenidos por LSI pero a la vez alivien sus desventajas. Por ejemplo, una elección natural puede ser un algoritmo basado en reglas de asociación, el cual detecta correlaciones o co-ocurrencias de términos, de manera eficiente. En términos prácticos, reglas con 2 o 3 términos podrían ser suficientes. Además, las reglas son fáciles de comprender.

Plan General (etapas)

1. Marco teórico respecto a Web search, LSI, SVD, Algoritmos de ML.
2. Revisión del estado del arte: recuperación de información en la web
3. Diseño de nuevo enfoque basado en ML para recuperación de información
4. Benchmarking de LSI vs. Método de ML propuesto
5. Análisis de resultados.
6. Difusión de resultados (publicaciones).

Literatura relevante (2-3 referencias)

1. Deerwester et al., Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990.
2. http://lsi.research.telcordia.com/

Tutor(es):

Dr. Claudio Meneses (Computer Science) – cmeneses@ucn.cl