Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Composite Retrival: Algo

Tesis presentada para optar al título de
Licenciado en Ciencias de la Computación

Amit Stein, Juan Andrés Knebel

Director: Obi-Wan Kenobi

Codirector: Master Yoda

Buenos Aires, 2014

Índice general

1. INTRODUCCIÓN

1.1. Descripción

Partiendo del tema planteado en el paper *"Composite Retrieval of Diverse and Complementary Bundles"* en el cuál se buscan elementos complementarios que persiguen un mismo objetivo.

1.2. Primeros pasos

Un gran poder conlleva una gran responsabilidad.
Tío Ben.

Se utilizó la base de datos de *"A Data-Driven Journey through Software Engineering Research"* con el objetivo de obtener conjunto de bundles complementarios. La base de datos contiene información sobre cerca de 5000 papers y sus 7000 autores, además cuenta con un perfil para cada paper, llamado `topicProfile` que nos ofrece una clasificación en porcentajes de que tema se refiere cada paper. Los temas incluidos son:

- Adaptive Systems
- Algorithm
- Architectures
- Artificial Intelligence
- Autonomic Systems
- Concurrency
- Database
- Distributed Systems
- Education
- Embedded
- Empirical Software Engineering
- Evolution
- Formal Methods
- Hardware

- Human Computer Interaction
- Information Systms
- Knowledge Engineering
- Languages
- Maintenance
- Methodologies
- Metrics
- Models
- Operating Systems
- Performance
- Process
- Product Lines
- Program Analysis
- Program Comprehension
- Real Time Systems
- Reliability
- Requirements
- Reverse Engineering
- Security
- Services
- Software Quality
- Synthesis
- Testing
- Visualization

Para generar la soluciones de bundles se debe definir algún atributo de los items para comparar su similitud y otro para conocer su complementariedad. Además para la búsqueda de soluciones se define una variable γ $(0 < \gamma < 1)$ usada para ponderar la selección de bundles y de esta manera obtener soluciones más cohesivas en las que cada bundle esta conformado por items muy parecidos entre sí o soluciones con bundles más dispersos en dónde es más importante que la relación entre los bundles sea lo más alejada posible
. A partir de información provista por la base de datos y con la definición de funciones de similitud, que se verán más adelante, tanto para autores y papers se generaron conjuntos de bundles para los siguientes criterios:

- Papers de tópicos similares que se presentaron en conferencias de distintos lugares.
- Autores similares de distintos lugares.

2. DESARROLLO

2.1. Definición de similitud y complementariedad

Como paso preliminar a la búsqueda de las soluciones se definió la noción de similitud y complementariedad para los grupos de elementos.

2.1.1. Papers

Ya se contaba con los perfiles de los papers, con lo cuál decidimos definir la similitud entre papers a través de la distancia entre los perfiles de cada uno de ellos. Cada perfil se interpreta como un vector de `n` posiciones, dónde cada posición pertenece a un tópico particular. Entonces la similitud entre dos papers la interpretamos como el ángulo vectorial entre dos vectores.
La complementariedad de dos papers se definió al lugar de presentación de dicho paper.

2.1.2. Autores

Para determinar que "tan iguales"son dos autores tuvimos que crear primero un perfil de autor para luego poder compararlos. Para lograrlo lo primero que se hizo fue tener en cuenta todos los perfiles de papers en los que participaron cada uno de ellos. Como cada paper tenía ya calculado su perfil el cuál representamos como un vector de `n` posiciones, dónde cada posición pertenece a un tópico particular. De esta manera tomamos todos perfiles (vectores) de los papers en los cuáles el autor participó e hicimos una suma componente a componente para así obtener un perfil de cada uno de los autores.
El paso siguiente fue determinar la similitud de los autores, la primer aproximación fue similar al cálculo de similitud de los papers y calculamos los ángulos entre todos los vectores. Como similitud alternativa decidimos restar el perfil de cada autor componente a componente y calcular el valor de su norma, así de esta forma a mayor norma mayor similitud entre autores.
La complementariedad de dos papers se definió al lugar de pertenencia del autor en cuestión.

2.2. Generación de soluciones

2.2.1. Papers

Originalmente la base de datos contenía unos 7777 papers, de los cuáles se tuvo que hacer una depuración, ya que había papers que no tenían ningún autor asociado o perfil creado. Luego de la depuración obtuvimos 4937 que cumplen los requisitos para la búsqueda de las soluciones.
Se generaron soluciones con distintos valores de γ $(0,1;\ 0,3;\ 0,5;\ 0,7;\ 0,9)$, las cuáles fueron de 10 bundles con 5 items cada una.
Las heurísticas utilizadas fueran la generación jerárquica (**HAC**) y la aleatoria (**BOBO**) para la producción de bundles y una estrategia en forma de algoritmo goloso para la selección de los bundles.

3. RESULTADOS

4. CONCLUSIONES