

Preface

A lo largo de este curso de Sistemas de Información, el cual se llevó acabo con éxito en la modalidad de enseñanza *a distancia*, hemos adquirido los conocimientos necesarios para enfrentarlos a los problemas y proyectos de la vida real que involucran las problemáticas de los Sistemas de recuperación de Información. Después de todo este tiempo estamos en condiciones de diseñar e implementar un **SRI** como parte de la evaluación final de esta asignatura.

Table of Contents

Proyecto Final de Sistemas de Información

Proyecto Final de Sistemas de Información	1
<i>Daniel García, Denis Gómez</i>	

Proyecto Final de Sistemas de Información

Daniel Alberto García Pérez and Denis Gómez Cruz

Universidad de la Habana, C-512

Abstract. Este documento abarcará de forma breve los aspectos más importantes relacionados con el proceso de diseño, implementación y análisis de los resultados del sistema. El Sistema de Información a obtener es el resultado de la realización del proyecto final de la asignatura Sistemas de Información.

Keywords: modelo vectorial, indexado, precisión, recobrado, retroalimentación

1 Principales características del problema

El sistema a desarrollar debe comprender todas las etapas del proceso de recuperación de información. Es decir, desde el procesamiento de la consulta hecha por un usuario, la representación de los documentos y la consulta, el funcionamiento del motor de búsqueda y la obtención de los resultados. No hay limitaciones respecto al Modelo de Recuperación de Información que deben emplear, puede ser cualquiera de los clásicos o alguno de los alternativos, siempre atendiendo a las características de cada uno y su adecuación al escenario en el que se aplicarán.

La evaluación del sistema debe realizarse usando las métricas objetivas y subjetivas estudiadas en clase empleando al menos dos colecciones de prueba distintas, incorporando consultas de ejemplo con los resultados proveídos por la solución, al menos una por cada colección de prueba.

2 Modelación

En la disyuntiva de sobre cuál modelo de recuperación de información (**MRI**) usar decidimos en primer lugar tomar alguno de los tres modelos clásicos: *modelo booleano*, *modelo vectorial* y *modelo probabilístico*. Dichos modelos se explicaban muy al detalle y tienen dedicadas dos conferencias del curso y se complementaba muy bien con la bibliografía recomendada en las mismas conferencias. Más aún se nos explicó las ventajas y desventajas de cada modelo y los escenarios dónde es más conveniente usar cada uno.

Descartamos en primer lugar el *modelo booleano* sobretodo porque sabemos que no establece un *ranking* entre los documentos recuperados y esto iba un poco en contra de nuestra idea de la aplicación visual final que queríamos que tuviera las mecánicas más frecuentes hoy en día como **Bing** o **Google**; además de que

nos limitaba el formato en que se debía escribir la consulta, siendo una forma para especialistas y no para usuarios en general como queríamos para nuestra aplicación.

También descartamos el *modelo probabilístico*, en este caso si obtendríamos las respuestas en forma de *ranking*, pero este modelo tiene una condicionante para su implementación como es la necesidad de tener un conjunto de documentos relevantes inicial, además de que la frecuencia de los términos en los documentos y consultas es irrelevante. Finalmente decidimos usar el *modelo vectorial*.

2.1 Modelo Vectorial

Teniendo en cuenta las ventajas y características de este modelo, fue el que finalmente usamos para modelar nuestro sistema. Buen rendimiento en la recuperación, coincidencia parcial entre consultas y documentos recuperados, y como punto más importante la obtención de los documentos recuperados en un orden de similitud con la consulta. Además de que es relativamente simple de implementar y de poder agregarle la capacidad de retroalimentación.

Sin entrar mucho en detalles de este modelo, sabemos que después de la fase de indexado que explicaremos más adelante en la sección de implementación, teniendo por cada documento un vector de frecuencias de términos; está una segunda fase donde se construyen los vectores de pesos que usa este modelo para cada documento. Una vez el sistema está funcionando este proceso se repite con el texto de la consulta (como si se tratara de un documento) insertada por el usuario, y finalmente se ordenan los documentos usando la función de similitud entre el vector de peso de la consulta y cada uno de los vectores de los documentos, siendo esta ordenación la recuperación que se le muestra al usuario.

2.2 Retroalimentación

Para la retroalimentación usamos el *algoritmo de Rocchio* estudiado en conferencias. La funcionalidad del mismo se adecúa muy bien al diseño visual de nuestra aplicación: el usuario, después de visualizar los resultados de su consulta puede elegir cuáles de los documentos recuperados son relevantes para él, luego usando el algoritmo mencionado, se usa esta información para construir una nueva consulta que se aproxima más a los documentos seleccionados y mucho menos a los que no se seleccionaron, y se vuelven a mostrar los resultados de la nueva consulta.

3 Implementación

El sistema fue implementado usando el lenguaje **Python**. Se crearon dos módulos, uno (`./models/`) donde se implementó la clase que representa el modelo vectorial (`./models/vectorial.py`) y otro (`./indexer/`) para las herramientas utilizadas en el indexado (`./indexer/indexer.py`).

La aplicación visual, que se trata de una aplicación web fue implementada en `server.py`, la cual corre un servidor accesible desde la red por vía de un navegador y muestra una interfaz para realizar las consultas. En el archivo `README.md` se especifican correctamente los pasos para correr el servidor y como usar la colección de preferencia. En el archivo `main.py` se implementó una versión de consola del SRI muy simple y fácil de usar en modo interactivo.

De igual forma se implementó un *script* para la visualización de la evaluación del SRI usando las métricas sugeridas en la orientación, se trata de `eval.py` y el mismo usa los *datasets* que se encuentran en la carpeta `./datasets/`. Se muestran los resultados consulta por consulta de prueba, mostrando un desglose de cada métrica. Al final de la ejecución se muestran resultados generales de la evaluación. De igual forma en `README.md` se encuentran las indicaciones para usar adecuadamente este *script*.

3.1 Indexado

Para el indexado(`./indexer/indexer.py`) se usó una herramienta previamente implementada como librería de Python, se trata de `yake` y se usa para la extracción de palabras o términos claves de un texto(*keyword-extraction*).

Iterando por cada documento de la colección seleccionada, se van extrayendo estos términos claves y se crea finalmente un conjunto general de todos los términos, el que representa las componentes de nuestros vectores. Luego usando el algoritmo de Aho-Corasick implementado en `./indexer/aho_corasick.py` se obtienen las frecuencias de cada término en cada uno de los documentos y se generan los vectores de frecuencias, dicha información se almacena en forma de `json` para su posterior uso por el modelo vectorial. Todo este proceso está encapsulado en el *script* `load_vectors.py`.

De manera análoga se procede con el texto de la consulta, usando los términos claves previamente calculados en el proceso anterior y usando nuevamente el algoritmo de Aho-Corasick se calcula el vector de frecuencias de la consulta, con dicho vector ya se pueden obtener los resultados usando el modelo implementado.

3.2 Parámetros del modelo usado

A la hora de calcular el vector ponderado de la consulta se usa un factor $0 \leq \alpha < 1$ de suavizado, en nuestra implementación este factor por defecto toma valor 0.5, pero puede ser fácilmente modificado si que desea en un futuro.

De igual forma ocurre con los parámetros presentes en el cálculo de la consulta óptima de la retroalimentación: α, β y γ que por defecto toman los valores: $\alpha = 1, \beta = 0.75, \gamma = 0.15$, tal cual se sugiere en la conferencia.

En cuanto a la relación del apartado visual con los documentos recuperados que se muestran, destacamos que siempre mostramos los 10 primeros resultados, aunque el resto se encuentran paginados en una barra al final de la lista de resultados.

4 Evaluación

Para la evaluación usaremos las medidas sugeridas en conferencia que son: **precisión**, **recobrado**, **medida F** y **medida F1**. Recordemos que en el caso del modelo vectorial no se recupera un conjunto en concreto de documentos, sino que estos se ordenan por relevancia, luego para que estas medidas sean calculables debemos seleccionar donde truncar la ordenación y tomar los primeros k elementos y considerar estos como el conjunto de documentos recuperados.

Usaremos los *datasets* **CRAN** y **CISI** para mostrar los resultados.

Empecemos por **CRAN**, tomando los $k = 10$ primeros documentos, he aquí una consulta de ejemplo evaluada:

- Consulta: `what design factors can be used to control lift-drag ratios at mach numbers above 5 .`
- Precisión: 0.4
- Recobrado: 0.16
- Medida F (Beta= 1.5): 0.19622641509433963
- Medida F1: 0.22857142857142856
- Fallout: 0.004369992716678805

Resultados finales después de analizar todas las consultas:

- Total de consultas: 225
- Precisión media: 0.295
- Recobrado medio: 0.424
- Fallos: 17/225 (7.556%)

Resultados finales, en este caso con $k = 20$:

- Total de consultas: 225
- Precision media: 0.195
- Recobrado medio: 0.537
- Fallos: 11/225 (4.889%)

Continuemos de forma análoga con **CISI**, tomando los $k = 10$ primeros documentos:

- Consulta: `An automated document clustering procedure is described which does not require the use of an inter-document similarity matrix and which is independent of the order in which the documents are processed. The procedure makes use of an initial set of clusters which is derived from certain of the terms in the indexing vocabulary used to characterise the documents in the file. The retrieval effectiveness obtained using the clustered file is compared with that obtained from serial searching and from use of the single-linkage clustering method.`
- Precisión: 0.3
- Recobrado: 0.5

- Medida F (Beta= 1.5): 0.41489361702127664
- Medida F1: 0.37499999999999994
- Fallout: 0.004814305364511692

Resultados finales después de analizar todas las consultas:

- Total de consultas: 76
- Precisión media: 0.339
- Recobrado medio: 0.133
- Fallos: 8/76 (10.526%)

Resultados finales, en este caso con $k = 20$:

- Total de consultas: 76
- Precisión media: 0.274
- Recobrado medio: 0.198
- Fallos: 5/76 (6.579%)

5 Análisis


Después de realizadas estas evaluaciones podemos sacar varias conclusiones: la precisión del **SRI** no es de la mejores pero está muy aceptable, rondando el 20% y 30%, lo que indica que 1 de cada 5 de los documentos recuperados es relevante para el usuario. Podemos apreciar también que tiende a disminuir si aumentamos la cantidad de documentos en el *top*.


Por otro lado tenemos el recobrado, que en el caso de **CRAN** es bueno, casi un 50%, sin embargo con **CISI** disminuye considerablemente al 15%. Y como es lógico, se puede notar en ambos *datasets* que al aumentar el *top* mejora el recobrado, dado que hay más posibilidades de que se contengan documentos relevantes en el *top*.

Finalmente, las veces que falló, que se refiere a cuando no se recuperó ningún documento relevante o lo que es lo mismo, cuando la precisión y el recobrado son nulos. Podemos apreciar que tienden a bajar del 10%, y claro está, esto mejora en la medida que aumentamos el *top*.

Pasemos ahora a los aspectos subjetivos. La aplicación web tiene una interfaz simple y muy fácil de usar, es intuitiva, con poner la consulta y dar *enter* se muestran casi instantáneamente los resultados. Los resultados se muestran paginados y permite al usuario rápidamente explorar un gran número de documentos. Se tiene una vista lateral donde se visualiza el contenido del documento que se seleccione con un *click*.

El código fuente de la implementación se puede encontrar [aquí](#).





some problems on heat conduction in stratiform bodies .
some problems on **heat conduction** in stratiform bodies . problems on **heat conduction** in multilayer bodies lead usually to complicated calculations . the present paper gives an idea of specific difficulties arising in the case of infinite composite solids . general deductions are applied to a specia ...

laminar hypersonic trail in the expansion-conduction region .
laminar hypersonic trail in the expansion-conduction region . the usual procedure in calculating the cooling process in a wake behind a blunt object is to assume a region of pure expansion up to a distance where the pressure has reached its ambient value, followed by a region where the mechanism of ...

one-dimensional transient heat conduction into a double-layer slab subjected to a linear heat input for a small time interval .
one-dimensional transient **heat conduction** into a double-layer slab subjected to a linear **heat** input for a small time interval . analytic solutions are presented for the transient **heat conduction** in composite slabs exposed at one surface to a triangular **heat** rate . this type of heating rate may occur ...

thermal analysis of stagnation regions with emphasis on heat-sustaining nose shapes at hypersonic speeds .
thermal analysis of stagnation regions with emphasis on heat-sustaining nose shapes at hypersonic speeds . the leading edges and noses of hypersonic vehicles are subjected to severe aerodynamic heating and must be cooled in some manner—such as, internal convection, transpiration, or radiation ...

conduction of fluctuating heat flow in a wall consisting of many layers .
conduction of fluctuating **heat** flow in a wall consisting of many layers . van gorcum has pointed to interesting and important analogies between the theory of a passive four-pole and the **conduction** of **heat** waves through stratiform bodies . this paper generalizes in certain regards van gorcum's ideas ...

a practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type .
a practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type . three approximate methods for the solution of the nonlinear equation of **heat** flow in a medium where **heat** is being generated by a chemical reaction are compared . the equations are ...

one dimensional heat conduction through the skin of a vehicle upon entering a planetary atmosphere at constant velocity and entry angle .
one dimensional **heat conduction** through the skin of a vehicle upon entering a planetary atmosphere at constant velocity and entry angle . closed-form solutions of the one-dimensional heat-conduction equations for the flow of **heat** into a plate with a laminar boundary layer have been obtained for a co ...

conduction of heat in composite slabs .
conduction of **heat** in composite slabs . a method of calculating the total quantity of **heat** that passes through a unit area from zero time to time t is developed . allowance is made for surface resistance by regarding each contact resistance as an additional layer of the appropriate thermal resistanc ...

an approximate treatment of unsteady heat conduction in semi-infinite solids with variable thermal properties .
an approximate treatment of unsteady **heat conduction** in semi-infinite solids with variable thermal properties . this very short paper presents an approximate procedure for the calculation of unsteady **heat conduction** in semi-infinite solids with variable thermal properties . it is claimed to be an i ...

conduction of heat in a solid with a power law of heat transfer at its surface .
conduction of **heat** in a solid with a power law of **heat** transfer at its surface . the nonlinear boundary value problem, where m and n are constants, is solved formally by first introducing power series in t for the unknown temperature and flux at the surface and then determining the coefficients in the ...

1

2

3

4

5

6

7

8

9

10

© 2021 D&D | [About](#)

Fig. 1. Ejemplo de la aplicación web

References

1. Carlos Fleitas: Conferencias de SI, Curso 2020-2021. Conferencias 2 y 4.