Medir cercanía entre documentos

Si se hace de una cantidad de documentos un index de todas las palabras usadas, se estabelece un espácio de palabras, si usamos para cada palabra una dimensión. Este espácio de palabras podemos llamar **I**, como index.

Cada documento se puede proyectar a este espácio de palabras, si es necesario, añadiendo dimensiones para palabras nuevas. Esto sería *indexar* este documento. Introduciendo un nuevo documento se genera un vector $\vec{d}$ para este documento así:

Contamos para cada palabra la cantidad de su utilización en este documento, utilizando el mismo orden como en **I**, ya está listo el vector $\vec{d}$ de este documento. Típicamente $\vec{d}$ sería una lista de números, incluyendo muchos zeros (sería fácil comprimir estos vectores también).

¿Cómo podemos aprovechar de este orden?

Me pregunté, qué nos daría el producto escalar ("dot product" en inglés) de vectores de dos documentos del mismo index. Claro que preparar un index es casí realizar también un motor de búsqueda, faltaría solamente hacerlo robusto de acentos y permitir el manejo de partes de palabras.

Si calculamos el producto escalar[1] c entre $\vec{d_1}$ y $\vec{d_2}$:

$$c = \vec{d_1} \vec{d_2}, \tag{1}$$

se multiplicarían similitudes entre los espácios de palabras usadas. Entonces, en el ejemplo de un miembro de MISTICA, se podría ofrecer automáticamente documentos, que con probabilidad alta pueda interesar este miembro. Había que ver, si esto sería un método para categorizar generalmente grandes cantidades de documentos.

Tomar en cuenta palabras de uso frecuente

Para precizar el simple producto escalar, podemos generar un vector común $\vec{p}$, que sirve para filtrar palabras de mucho uso, como el producto está fuertemente influído por palabras no muy caracteristicas. Con este

[1] El producto escalar se calcula por ejemplo para dos vectores $\vec{a} = (a_1, a_2), \vec{b} = (b_1, b_2)$ de sólo dos dimensiones así: $c = a_1 b_1 + a_2 b_2$

vector de pesos o probabilidades podemos multiplicar los vectores de documentos, para producir mejor calidad. El producto escalar se calcularía entonces:

$$c = \sum_{i=1}^{dim(\mathbf{I})} d_{1_i} p_i^2 d_{2_i} \tag{2}$$

¿Cómo calculamos $\vec{p}$?

Una posibilidad sería tomar lo inverso del uso de la palabra i. Esto no tomaría en cuenta la existencia de áreas temáticas:

$$\vec{p} = p_i = \frac{\text{Cantidad de todas las palabras}}{\sum_{d=1}^{\text{Todos los documentos}} \text{Cantidad de palabra } i} \tag{3}$$

Otra posibilidad, que parece mas segura, pero seguramente mas limitada, sería definir palabras i que parecen en cada documento como zero en p_i, el resto como uno. Esto sería la manera trivial, que no funcionará bien, si hayan muchos documentos cortos.

Implementación

Los algoritmos para implementar esta sensitividad de contexto son fáciles, y también suficientemente rápidos, pesar de que crece en quadrado la cantidad de calculaciones.

En el mismo tiempo se puede realizar el motor de búsqueda, que podría permitir muestrar "documentos temáticamente cercanas".

Supongo que sería un día de trabajo realizar esto.

Sería útil indexar las páginas del metasitio también para aprovechar de este mecanísmo. Los diferentes idiomas había que indexar separados, para que no se mezclarían semanticas diferentes.