Expansión semántica para mejorar la diversidad en la formulación de consultas

Elliot Ide-Pozo

Instituto de Informática, Universidad Austral de Chile, Valdivia, Chile
elliot.ide@alumnos.uach.cl
http://www.inf.uach.cl/

Resumen Aunque la diversidad de los resultados ha sido estudiada desde los primeros sistemas de recuperación de información, existen pocos estudios que exploren la diversidad y su representación en un contexto educacional. Inherentemente los enfoques que buscan apoyar las dificultades en la búsqueda web, tales como la sugerencia y la expansión de consultas, se enfocan en maximizar la relevancia de los resultados de la consulta original. En este trabajo se presenta un método que integra relaciones semánticas mediante Word Embedding para la expansión con retroalimentación ciega. Utilizando un corpus basado en los registros de consultas de estudiantes, se entrenan 3 modelos Word2vec para obtener términos semánticamente relevantes a cada consulta. Se estudia la arquitectura propuesta en una tarea de búsqueda puntual, acotando el número de términos candidatos en cada modelo según la frecuencia mínima de palabras. Finalmente se compara la diversidad en dos grupos de consultas, midiendo la similitud léxica de los snippets de los resultados antes y después de la expansión. Los resultados indican la posibilidad de mejorar la diversidad mostrando además que una menor similitud semántica puede conducir a una mayor diversidad.

Abstract Although the diversity of results has been studied since the first information retrieval systems, few studies explore diversity and its representation in an educational context. Inherently, approaches that seek to support difficulties in web search, such as query suggestion and query expansion, are focused on maximizing the relevance of results over the original query. This work presents a method that integrates semantic relationships using Word Embedding for expansion with blind feedback. Using a corpus based on the user's query logs, three Word2vec models are trained to obtain semantically relevant terms for each query. The proposed architecture is studied in a specific search task, limiting the number of candidate terms in each model according to the allowed frequency of words. Finally, the diversity in two groups of queries is compared, measuring the lexical similarity of the snippets of the results before and after the expansion. Results indicate the potential for improving diversity, also showing that lower semantic similarity can lead to better diversity.

Keywords: Expansión de consultas · Diversidad · Similitud semántica · Word2vec · Retroalimentación ciega · Aprendizaje basado en búsquedas

1. Introducción

Pese a que las nuevas generaciones tienen una mayor y temprana exposición a la tecnología, la búsqueda de información como actividad online puede resultar desafiante, principalmente al momento de formular una consulta o query [10].

Ante la expresión de esta necesidad, los motores de búsqueda o sistemas de recuperación de información (IR) están diseñados con un propósito general y no consideran necesariamente un contexto educacional o una intención de aprendizaje. Por otra parte, estudios han mostrado que existen claras diferencias entre adultos y jóvenes en las necesidades de información, enfoques de búsqueda y habilidades cognitivas [9].

Investigaciones anteriores han mostrado además que un gran porcentaje de las consultas emitidas son repeticiones o reformulaciones [3]. Jiang y Ni (2016) mostraron que cerca de un 65 % de las consultas de los usuarios son reformulaciones.

Independientemente, existe una tendencia hacia la elaboración de consultas ambiguas y de baja longitud, lo cual impacta en la *diversidad de los resultados* [15], y en consecuencia podrian guiar a una visión parcial y sesgada de la necesidad de información.

Ante esto, se han desarrollado múltiples métodos para apoyar la formulación, por lo regular, modificando la consulta emitida. En esta investigación, se abordan dos enfoques ampliamente explorados: 1) Expansión de consultas o Query Expansion (QE). 2) Sugerencias de consultas o Query Suggestion (QS).

El primer enfoque propone la incorporación de términos significativos para remover la ambigüedad natural del lenguaje, permitiendo expresar de forma más detallada el concepto de información *en la consulta original* [23]. Por otro lado, QS hace énfasis en predecir el objetivo de búsqueda del usuario, en base a su *comportamiento de búsqueda*, sugiriendo nuevas consultas [11].

Pese a sus diferencias, estos enfoques comparten similitud en las técnicas de modificación de consultas que utilizan. En particular, en esta investigación se hace énfasis en la integración de relaciones semánticas, para un mayor entendimiento de las intenciones del usuario [28].

Por otra parte, dentro de la literatura de QE y QS, existen pocos algoritmos que consideren la diversidad como factor en un contexto educativo. La mayoría de estos métodos están orientados a maximizar la relevancia de los resultados o documentos obtenidos, en función de la consulta inicial [15].

Sobre esto, una efectiva expansión puede mejorar los resultados de búsqueda y por tanto ayudar al cumplimiento de los objetivos de información [32].

En esta investigación, se estudia cómo extraer términos que aumenten cuantitativamente la diversidad de los resultados sobre una plataforma de aprendizaje basado en búsquedas.

El desafío implica apoyar la formulación de la consulta asumiendo una intención parcial de aprendizaje por parte del usuario, y en el proceso conducir mediante *expansión* a otros matices del tópico en cuestión.

Partiendo de una definición tradicional de diversidad [7], se establece una arquitectura para la obtención de términos, utilizando modelos Word2vec basados en Skip-Gram. El enfoque consiste en explotar la *representación vectorial* (embedding) altamente semántica del vocabulario del modelo [20], empleando como colección de entrenamiento, un corpus seccionado de resultados de consultas relacionadas a la tarea de búsqueda.

La diversidad es evaluada en los nuevos resultados, utilizando una métrica propia de *disimilitud léxica*. Para esto, la expansión es realizada en dos subconjuntos, determinando los candidatos según su *similitud semántica* con el último término de cada consulta. Esta separación se hace con el fin de explorar la diversidad en un plano de reformulación de consultas.

Esta propuesta se ubica en el enfoque de QE como retroalimentación ciega de relevancia o PRF, incorporando QS con el uso de los registros de consultas. Los resultados indican la posibilidad de mejorar la diversidad estimada, mostrando además que una menor similitud semántica puede conducir a una mayor diversidad.

Se abordan las siguientes preguntas de investigación:

- RQ-1: ¿La expansión de consultas basada en semánticas puede aumentar la diversidad de los resultados?.
- RQ-2: ¿Cómo se relacionan la similitud semántica y la diversidad estimada?.

2. Trabajo Relacionado

A continuación se repasan las investigaciones más relevantes para este proyecto, empezando con el concepto de relevancia y un breve resumen del enfoque sobre el cual se basa la propuesta. Luego, se discute el manejo de diversidad en la literatura, haciendo mención a dos perspectivas para su incorporación, y el uso de fuentes de información para la obtención de términos. Finalmente, las dos últimas secciones comprenden la elección de la métrica de similitud textual para la comparación de diversidad y la formulación del problema.

2.1. Relevancia y expansión de consultas

La incerteza de las consultas es un problema que existe ampliamente en el escenario de la búsqueda web [15] y que afecta en la capacidad de recuperar documentos relevantes por el sistema IR [30], también conocida como efectividad de recuperación.

Conceptualmente, la relevancia es un factor crucial en la búsqueda de información. Bajo una perspectiva de usuario, se quiere presentar dentro de una colección de documentos, un número finito de estos que sea atingente a su consulta, con el propósito de satisfacer su necesidad de información en el menor tiempo posible.

Ante este desafío, técnicas como la expansión de consultas (QE), sugerencias de consultas (QS) y el autocompletar (QAC) han sido desarrolladas [23]. Es más, algunas de estas forman ya parte de buscadores como Google y Bing, aunque se desconoce propiamente tal el nivel de diversidad que consideran.

En particular, esta propuesta está basada en la expansión de consultas como método principal. Dentro de esta área, la expansión a utilizar se conoce como retroalimentación ciega de relevancia o pseudo-retroalimentación (PRF).

Este tipo de expansión tiene la mayor efectividad en la literatura [6], y se basa en la suposición de que los documentos mejor clasificados son relevantes para el documento deseado por el usuario. En otras palabras, los términos que los componen y su distribución pueden ayudar a descartar documentos irrelevantes de la colección de documentos.

Cabe mencionar que el acto de expandir una consulta, forma parte de las estrategias de reformulación de los usuarios para acotar una búsqueda [3].

Esta investigación busca definir un algoritmo que entregue términos que mejoren una diversidad estimada utilizando el buscador Bing. En un entorno donde no se tiene control sobre el sistema IR, la efectividad del método depende de la calidad de los resultados que este provee.

Aunque el foco de esta investigación es aumentar la diversidad de los resultados mediante la expansión de las consultas, la relevancia es un factor que no se puede apartar, en especial en el proceso de selección de términos. Se deben considerar características que vayan más allá de la frecuencia, para obtener términos que afecten positivamente la efectividad de la recuperación [6].

2.2. Diversidad y fuentes de información

Sobre el concepto y desafío de lidiar con la diversidad se pueden distinguir dos perspectivas principales:

1. La selección ajustada de resultados diversos, asociada al área de Búsqueda y Recuperación de Información (IR) referida como Search Result Diversification (SRD).
2. La sugerencia de consultas o términos que permitan diversidad en los resultados. Aquí se encuentra la expansión de consultas (QE), sugerencia de consultas (QS) y Query Auto-Completion (QAC).

En la primera perspectiva, los métodos SRD se separan en *explícitos* e *implícitos* [2]. En el primer caso se tiene una noción de la intención de búsqueda (intent-aware), por lo que se busca modelar los aspectos o subtópicos detrás de la consulta y aumentar la diversidad en estos.

Hu et al. [13] estudian la diversidad bajo la propuesta de una estructura jerárquica que representa las intenciones de los usuarios.

Al contrario, los métodos implícitos lidian con el desconocimiento basándose en la suposición de que una mayor disimilitud entre documentos satisface un mayor número de intenciones. Carbonell y Goldstein (1998) proponen uno de los primeros algoritmos implícitos de diversificación, estableciendo una métrica que combina la relevancia de una consulta con la novedad de información, aplicando un reordenamiento de documentos basado en diversidad [7].

En la segunda perspectiva, que es el *principal* foco de interés, se da soporte al usuario en la entrega de la consulta al sistema. A diferencia de QE y QAC, para QS se entrega al usuario una serie de consultas que son semánticamente relevantes *posterior* a la emisión de la consulta [29].

Las investigaciones revisadas que siguen esta línea, abarcan en mayor grado la medición de diversidad resultante en su propuesta, asociada al reordenamiento de documentos, por sobre soportar la diversidad en la elaboración de la consulta.

Se detalla que esta investigación no considera evaluaciones de relevancia o diversidad del sistema IR, debido a que utiliza un sistema externo, por lo que el reordenamiento no forma parte del alcance.

Madrazo et al. [18] proveen recomendaciones considerando explícitamente las necesidades de información ambiguas y diversas basándose en análisis de texto y rasgos textuales asociados a contenidos para niños. En su propuesta aseguran la diversidad entre las consultas recomendadas bajo una estrategia de similitud de contenido, excluyendo sugerencias que se refieren a un mismo tópico.

Câmara y Santos (2019) proponen estrategias para la recomendación de consultas diversas atravesando un registro de consultas semánticamente anotado [5].

Por otro lado, Umemoto et al. [31] apoyan el descubrimiento de aspectos en tareas que requieren de un mayor volumen de información para su resolución (intrínsecamente diversas). Mediante una interfaz QS, muestran cuantitativamente la cantidad de información importante pérdida con cada consulta.

Referente a QE, Bouchoucha et al. [4] presentan su enfoque como expansión diversificada de consultas (DQE), haciendo uso del tesauro ConceptNet[1] para seleccionar términos de expansión diversos, bajo su propia métrica inspirada en la Relevancia Marginal Máxima (MMR) [7].

El método propuesto combina un lineamiento tradicional del reordenamiento de resultados, en el cual la poca similitud entre documentos promueve la diversidad [7], junto a los principios de tendencias de diversificación de QS que lidian con la incerteza, intentando cubrir la mayor cantidad de facetas posibles [15]. En las siguientes secciones, se explicita el uso de representaciones vectoriales o Word Embedding para comparar documentos y capturar relaciones semánticas.

Por otro lado, otro factor a considerar es la *fuente de información* para la obtención de términos. Las investigaciones más recientes incorporan el uso de fuentes externas para la extracción de conceptos: Besbes et al. [2] utilizan el buscador *Pubmed* como fuente de terminología en el área de Medicina. Fails et al. [11] analizan contenido escrito para niños junto para generar sugerencias de interés. Por otro lado, Duarte y Hiemstra (2012) proponen el uso de etiquetas de redes sociales para sugerir consultas relacionadas a tópicos infantiles [10]. Por último, Shajalal et al. [27] utilizan las sugerencias y completaciones de múltiples buscadores para seleccionar términos basados en aspectos del tópico, realizando expansión mediante clustering.

En cuanto al volumen de información, en la expansión de consultas mediante Word Embedding, se han obtenido resultados similares en *relevancia* para los modelos Word2vec y GloVe[2], utilizando documentos restringidos por tema por sobre colecciones globales [8].

Tradicionalmente, los métodos QS utilizan los registros de las consultas o *query logs* para las recomendaciones [21], aunque esta información no suele estar siempre disponible por políticas de privacidad.

Este tipo de información se asocia en la literatura a una dimensión de *contexto* en la personalización del proceso de búsqueda y ayuda a identificar los intereses e intenciones del usuario [29].

Para este caso, se tiene la plataforma de aprendizaje como intermediario de la fuente de información que corresponde al buscador Bing, la cual almacena los registros de consultas incluyendo los resultados de cada consulta emitida.

2.3. Métricas de similitud sobre texto

Para evaluar si existe una mejora sobre la consulta original se requiere una métrica que estime la diversidad de los resultados. Como se mencionó anteriormente, se utilizará una métrica de similitud sobre texto, con la premisa de que un valor bajo de similitud implica una mayor diversidad.

Este tipo de métrica ha sido utilizada ampliamente en la clasificación de texto y algoritmos de clustering [17] y puede ser *léxica* o *semántica* dependiendo del origen de la similitud.

[1] https://conceptnet.io/
[2] https://nlp.stanford.edu/projects/glove/

Se habla de similitud léxica, cuando la comparación se basa únicamente en la secuencia de caracteres que componen las palabras. La similitud semántica en cambio, considera el significado y contexto en la estimación, e implica la adquisición de información de una fuente mayor [12].

En la comparación de documentos se emplea usualmente una representación vectorial, donde cada componente indica el valor de una característica del documento como el número de ocurrencias de un término.

Para comparar resultados entre sí, la estimación de características se realiza mediante la métrica *TF-IDF* (frecuencia de terminos/ frecuencia inversa de documentos) [17]. Luego, para estimar diversidad sobre los resultados se utiliza la metrica de *similitud del coseno*. Una métrica usada ampliamente en IR y estudios relacionados [1], independiente del tamaño del texto y que se ve afectada por la repetición de términos en su estimación [12]. En la sección 3.2 se explica con mayor detalle la medición de diversidad.

2.4. Formulación del problema

En la recomendación de consultas se identifican enfoques basados en: el *uso de grafos* para encontrar clusters de consultas, *redes neuronales profundas* para sintetizar sugerencias y en menor grado la *integración de semánticas* [14], como medio para una comprensión más profunda de las intenciones de la consulta [28].

El método propuesto incorpora la integración de semánticas en el proceso de expansión. Es *necesaria* una relación semántica entre la consulta y el término sugerido a fin de asegurar la relevancia de los resultados sobre la intención original del usuario.

En este sentido, las representaciones vectoriales mediante redes neuronales son de interés, debido a que codifican explícitamente múltiples patrones lingüísticos [20].

En particular, se utiliza la técnica de procesamiento natural de lenguaje (NLP) *Word2vec*, para obtener una representación del vocabulario del corpus que permita una comparación semántica.

Esta técnica puede representar las palabras en un espacio vectorial preservando su proximidad semántica [14]. Dicho de otro modo, los vectores semánticamente similares están cerca entre sí, de la misma forma que las palabras más cercanas en una oración tienen una mayor relación que las palabras distantes.

Se destacan las siguientes aplicaciones de Word2vec junto a QE: Kuzi et al. [16] utilizan PFR (relevancia ciega) operando en un modelo de lenguaje, y asignando puntajes a los términos sugeridos por el modelo Word2vec, aunque su énfasis es en mejorar la efectividad de la recuperación de información. Roy et al. [26] mediante Word2vec desarrollan una técnica QE donde los términos relacionados son obtenidos mediante un enfoque del k vecino más cercano.

Para la implementación del modelo existe la arquitectura Continuous Bag-of-Words (CBOW) y Skip-Gram. En ambos casos, se entrena una red neuronal de 1 capa oculta (sin función de activación), en la cual, los vectores de peso de la capa oculta codifican el significado de las palabras del corpus.

La diferencia entre estas arquitecturas, es el enfoque del entrenamiento. Para CBOW, el modelo aprende a predecir una palabra objetivo utilizando todas las palabras de su vecindario. En el caso de Skip-Gram el modelo predice múltiples palabras del contexto dada la palabra de entrada.

En términos de rendimiento, para un corpus del orden del billón de palabras, CBOW tarda hasta 3 veces menos que Skip-Gram, aunque este último tiene un mejor desempeño en evaluaciones de carácter semántico [19].

En relación a la expansión de consultas, Skip-Gram por definición provee una menor generalidad del contexto en su representación. El entrenamiento de esta arquitectura implica "observar" las palabras cercanas a una palabra de entrada en una ventana definida, entregando pares de palabras cerca del conjunto de frases del corpus. El resultado es una representación semánticamente relevante de todas las palabras del vocabulario considerando múltiples contextos.

Formalmente, se quiere hacer uso de este modelo para promover la diversidad en un entorno donde:

1. Se tiene una noción de la intención de búsqueda asociada a una tarea de búsqueda (información contextual).
2. Se cuenta con acceso a los registros de consultas.
3. No se tiene control sobre el sistema IR que provee los resultados.

Explícitamente, para una consulta q asociada a una tarea de búsqueda y a un conjunto de resultados $S = \{s_1, \ldots, s_N\}$, con un corpus C conformado por los registros de consultas previos, se provee una serie de términos $T = \{t_1, \ldots, t_M\}$ semánticamente relevantes, con los cuales se realiza la expansión de q o Q', obteniendo $m \times N$ resultados o S' cuya selección depende de su valor de diversidad en comparación a S.

3. Metodología

Bajo la arquitectura del método propuesto (Figura 1), se explora un concepto propio de diversidad sobre una tarea de búsqueda puntual. El proceso de extracción de términos candidatos y su posterior expansión surge a partir de una consulta y se resume en 5 etapas principales: 1) Preprocesamiento, 2) Entrenamiento del modelo, 3) Selección semántica, 4) Scraping y 5) Medición de diversidad posterior.

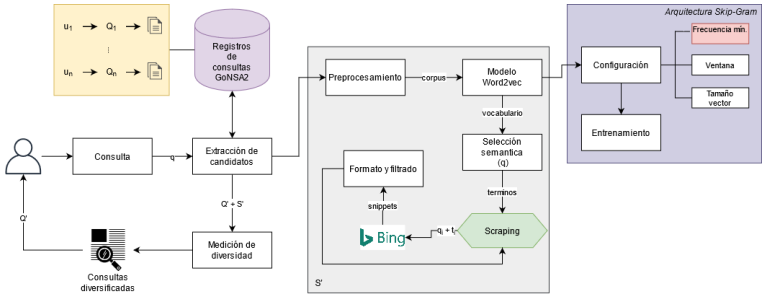

Figura 1. Arquitectura general del método

Como se mencionó en la sección 2.4, se cuenta con acceso a los registros de consultas de los usuarios. Para la aplicación del método se hace uso de una colección de resultados relacionados a la tarea de búsqueda, específicamente, los snippets de los resultados.

Particularmente, se estudia la *diversidad y similitud semántica* en dos grupos de consultas que se diferencian en su *contexto de emisión*. También se consideran 3 configuraciones del modelo Word2vec modificando el parámetro de frecuencia mínima de palabras. En las siguientes secciones se precisa con mayor detalle el uso del dataset, la métrica de diversidad definida y los análisis respectivos en torno al corpus para la definición de parámetros y su configuración.

3.1. Dataset

El conjunto de datos a evaluar proviene de los registros de GoNSA2, un marco de aprendizaje basado en búsquedas de información [22], del cual se utilizará únicamente la *información contextual* de la tarea de búsqueda, consultas emitidas mediante Bing y sus respectivos resultados. Cabe mencionar que las consultas fueron realizadas por 30 estudiantes de primer año de Ingeniería de la Universidad Austral de Chile de entre 18 y 19 años durante el año 2019.

Para cada consulta se tienen hasta 50 resultados almacenados junto a su título, url y *snippet*. Este último corresponde a texto que resume brevemente el contenido que se encuentra en el enlace. Dada la complejidad de extraer información relevante en páginas diferentemente estructuradas se utiliza el snippet como elemento textual de comparación.

En el Cuadro 1 se presenta información y estadísticas sobre el subconjunto de datos escogido.

Cuadro 1. Resumen y estadísticas de consultas

Tarea de búsqueda	**Objetivo**
Construir un automóvil	Identificar características necesarias para la construcción de un automóvil soapbox
Número total de consultas	345
Promedio de consultas por usuario	11.5 (DE = 10.59, MAX = 41)
Promedio de consultas únicas por usuario	5.3 (DE = 3.55, MAX = 15)
Promedio de términos por consulta	4.3 (DE = 2.10)

DE : Desviación Estándar, MAX : Valor máximo

Los snippets también serán utilizados en conjunto como corpus para entrenar a los modelos Word2vec, delimitando su volumen según un número de consultas máximo de todos los usuarios o *maxq*. En el proceso de generación del corpus se ignoran los snippets duplicados y las palabras vacías (stop words).

En promedio el rango de usuarios emitió 12 consultas, donde 5 de estas fueron únicas. Este último valor se utiliza como límite para estudiar el aumento de información del corpus en función del total de palabras únicas (PU) detectadas (Cuadro 2).

Cuadro 2. Descripción del corpus según el número de consultas máximo

maxq	**Palabras**	**Frases**	**Palabras únicas**	**Variación Porcentual (PU)**
1	27939	2798	8793	0 %
2	46597	4828	12875	46.42 %
3	58285	5991	15019	**16.65 %**
4	66483	6918	16179	7.72 %
5	75560	7908	17455	7.88 %
Resumen	54972.80 (18495.05)	5688.60 (1976.41)	14064.20 (3394.16)	

Considerando los snippets resultantes de la primera consulta de todos los usuarios, se tiene aproximadamente un 31 % de palabras únicas de un total de 27939 palabras. Sí se considera además, los snippets de la segunda consulta, el número de palabras únicas aumenta a 12875, teniendo una variación porcentual ($\frac{v_2 - v_1}{v_1} \cdot 100$; con v_2: *valor presente*, v_1: *valor pasado*) de un 46.42 % respecto al caso anterior.

Por otro lado, utilizando los resultados de hasta 5 consultas, se tienen cerca de 14064 palabras únicas. Para el entrenamiento de los modelos, se opta por utilizar un corpus constante considerando $maxq = 3$, cuyo número de palabras únicas es el más cercano al promedio, conteniendo cerca de un 26 % de palabras únicas del total de palabras.

3.2. Midiendo diversidad

Un grupo menor de las palabras en los snippets son únicas, en contraste, para $maxq = 3$, cerca de un 74 % son repeticiones, lo que sugiere la factibilidad de un enfoque léxico para medir disimilitud o diversidad.

Para comparar cuantitativamente sí existe una mejora en la diversidad, se establece una métrica basada en la *similitud del coseno* sobre un conjunto de N snippets.

Cada snippet s se compone de palabras que tienen un grado de relevancia (provisto por el sistema IR) a la consulta emitida q. Para medir la diversidad asociada a la consulta, esta se representa vectorialmente junto a un subconjunto de snippets aplicando una normalización *TF-IDF* [17].

El resultado son N vectores cuyos componentes indican la *relevancia* (en términos de frecuencia) de las palabras que los conforman inicialmente con respecto al vocabulario resultante, descartando de este las palabras vacías o stop words presentadas por la librería NLTK[3].

Posteriormente, se utiliza la métrica de similitud del coseno *sim* descrita en la ecuación 1, donde $\|s\|$ es la norma euclidiana.

$$sim(s_1, s_2) = \frac{s_1 \cdot s_2}{\|s_1\|\|s_2\|} \tag{1}$$

Esta métrica computa el ángulo entre un par de vectores entregando valores entre 0 y 1. Cuanto más se acerque el valor del coseno a 1, menor será el ángulo y mayor será la similitud entre los vectores.

Aplicando la función $f(x) = 1 - x$, la mayor diversidad entre snippets se encontrará opuestamente en los valores próximos a 1.

Finalmente, para una consulta q, con N snippets del conjunto asociado S, e $I = \{1, \ldots, N\}$, su valor de diversidad se define como la suma de similitudes de cada par único *sum_cos* (ecuación 2).

$$sum\_cos(q) = \sum_{\substack{i \in I \\ i<j}} \sum_{j \in I} sim(s_i, s_j), \quad s_i, s_j \in S \tag{2}$$

Así, un valor próximo a N implica una mayor diversidad *global* para la consulta, y su valor respectivo será un referente de qué tan diversos son los nuevos resultados para cada consulta incrementada q'.

3.3. Definición de parámetros

Para la obtención de términos candidatos se utiliza la implementación del modelo Word2vec de Radim Rehurek [24]. Entre sus múltiples parámetros se especifica: la dimensionalidad del vector *vector_size* 100 por defecto, el corpus *sentences*, la arquitectura *sg* que en este caso es Skip-Gram, la frecuencia mínima de palabras a considerar *min_count* y la ventana máxima window. El resto de parámetros permanece con su valor por defecto.

[3] Plataforma de Python para el procesamiento de lenguaje natural

Para la generación de los modelos a explicar a continuación, el corpus queda definido considerando:

- Hasta 3 consultas emitidas por usuario ($maxq = 3$).
- Uso de los resultados (snippets) de todo el conjunto de usuarios.
- Eliminación de snippets duplicados.

Junto a esto, la entrega de cada snippet se separa en las oraciones que lo componen. Es decir si un snippet contiene 3 frases, pasan a ser 3 elementos independientes dentro del corpus. Este punto se relaciona con el parámetro *window*, el cual permite fijar la distancia máxima entre la palabra de entrada (desde donde se aplica la ventana en la frase) y las palabras cercanas a considerar en el entrenamiento.

El vocabulario final que cada modelo acota está ligado a otros parámetros determinables, en especial, a la frecuencia mínima de palabras o *min_count*. Por otro lado, la evaluación de cada modelo depende del uso de las salidas, es decir, un modelo es mejor que otro si existe una mejora global en la diversidad incorporando las palabras cercanas estimadas.

A modo de respaldar las decisiones sobre los parámetros *min_count* y *window*, se estiman indicadores sobre el corpus basados en su definición (Cuadro 3).

Cuadro 3. Indicadores para elección de parámetros

Estadísticos del corpus para *maxq = 3*		
Total de palabras únicas	**Total de snippets**	**Total de frases**
15019	2325	5991
Descripción de indicadores		
Parámetro	**Indicador**	**Resultados**
min_count	Promedio de frecuencias	3.69 (DE = 11.34)
window	Promedio de palabras en frase	9.72 (DE = 6.93)
	Promedio de palabras en snippet	25.06 (DE = 6.71)

Sobre *min_count* se tiene un promedio de frecuencia de aproximadamente 4, para todas las palabras únicas a lo largo del corpus. En el Cuadro 4, se presenta una muestra aleatoria de palabras según su valor de frecuencia f en torno al promedio.

Cuadro 4. Muestra del corpus en torno al promedio de frecuencias

$f < 4$	$4 \leq f \leq 30$	$f >> 30$
modelado (3)	*plan* (30)	*auto* (473)*
estrategias (2)	*diseños* (21)	*construir* (323)
preparar (2)	*fabrica* (12)	*motor* (301)
pilas (2)	*procedimiento* (10)	*sistema* (106)
manufacturar (1)	*eficaz* (7)	*proceso* (82)

Cuantitativamente, cabe mencionar que las palabras con frecuencia 1 representan más de la mitad de las palabras únicas con un 56.6 % de contribución (Figura 2).

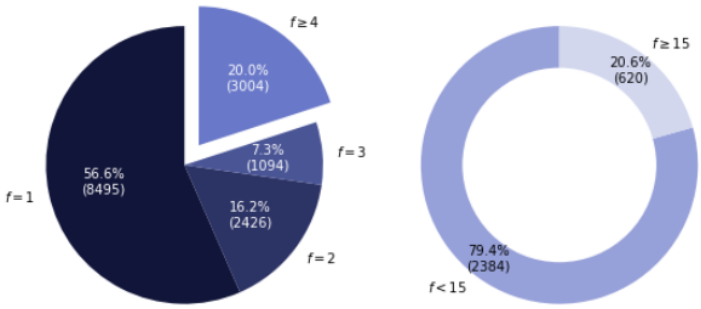

Figura 2. Distribución de frecuencias f de palabras únicas

Por otro lado, se cuenta también con información contextual de la tarea de búsqueda. Específicamente, se tienen 3 segmentos de texto que contienen una descripción de la tarea, un resumen y objetivo. Al intersectar las palabras únicas de este conjunto con el corpus, se tiene que un 80 % pertenece a este último (28 de 35 palabras).

En relación con este último punto, se definen 3 límites para acotar el vocabulario en cada modelo: un límite inferior de 5 (valor sugerido por defecto y próximo al promedio) que considera un 15.5 % del corpus, un límite intermedio de 8, para contener un 80 % de las palabras únicas más frecuentes de la información contextual, y un límite superior de 15 conteniendo un 4.12 % del corpus total.

Para el parámetro *window* se utiliza como indicador el promedio de palabras en frase. Cada frase está compuesta en promedio por al menos 10 palabras (9.72 , DE = 6.93). En este caso se utilizará para los 3 modelos el valor 5 por defecto del modelo.

3.4. Configuración y Scraping

Para la obtención de términos se entrenan 3 modelos con la misma arquitectura Skip-Gram y una ventana de 5 palabras:

- Modelo A o estándar: con un mínimo de frecuencia de 5. Genero un vocabulario de 2416 palabras.
- Modelo B: con un mínimo de frecuencia de 8. Genero un vocabulario de 1416 palabras.
- Modelo C: con un mínimo de frecuencia de 15. Genero un vocabulario de 662 palabras.

La reducción del vocabulario dada por *min_count* ocurre antes que el entrenamiento [24], por lo que la aplicación de la ventana es entre las palabras restantes de cada frase.

Esto hace que cada modelo pueda tener una representación diferente para una misma palabra, porque no se entrena con el mismo contexto, pese a usar la misma ventana. En consecuencia, también mientras más acotado esté el vocabulario, menor son las aplicaciones de cada palabra en otros contextos.

La selección de términos está basada en un método propio de la implementación de Radim Rehurek, el cual toma una o más palabras como entrada, retornando un conjunto de palabras ordenadas por su valor de similitud $similarity\_w$ definido como: *"la similitud de coseno entre el promedio de los vectores de peso de proyección de la palabra de entrada y los vectores de cada palabra del modelo"* [25].

Para utilizar este método, las palabras de entrada deben *pertenecer* al vocabulario que genera el modelo, además debe indicarse si sus vectores contribuyen aritméticamente de forma positiva o negativa a la estimación de similitud. Este último paso se refiere a la capacidad del modelo Word2vec de capturar relaciones semánticas en la representación vectorial, permitiendo observar relaciones del tipo: $rey - hombre + mujer \approx reina$.

En particular, se define la palabra de entrada a utilizar como el *último término de cada consulta original* que pertenezca al vocabulario.

Las consultas a analizar en términos de diversidad y similitud semántica, se dividen además en dos grupos, con el fin de explorar la diversidad en un plano de reformulación de consultas:

- Grupo 1 (G1): considera las primeras 12 consultas únicas emitidas por usuarios únicos, representando las consultas sin retroalimentación del buscador.
- Grupo 2 (G2): corresponde a las 7 consultas más emitidas por todos los usuarios, desde la segunda consulta. Se asocia al producto común del proceso de reformulación.

En promedio, las 19 consultas seleccionadas se componen de 4 términos (3.894) y tienen una diversidad promedio global de 2.867 (DE = 0.103). La Figura 3 describe mayores estadísticas sobre el conjunto en torno a su diversidad original y número de términos.

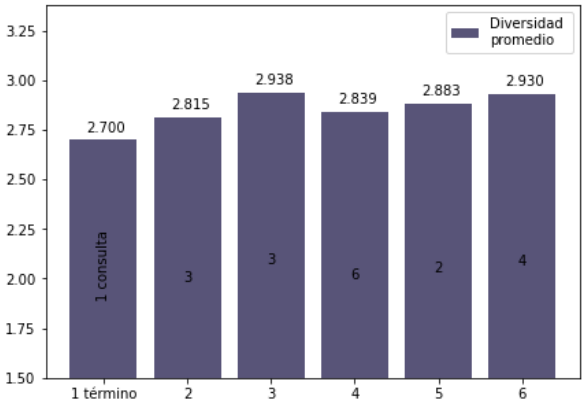

Figura 3. Diversidad promedio por número de términos (G1 y G2)

Para estudiar la diversidad posterior a la expansión, sobre cada consulta individual se obtienen las *50 palabras con mayor similitud por modelo* según su último término. Utilizando Selenium como herramienta para scraping, se emiten las 150 expansiones de cada consulta mediante el buscador Bing, adquiriendo en el proceso los 3 primeros snippets que no sean publicidad. Sobre un total de 2850 consultas realizadas, se procede a estimar la diversidad obtenida sobre los resultados para su análisis.

4. Resultados

Para responder la primera pregunta de investigación *¿La expansión de consultas basada en semánticas puede aumentar la diversidad de los resultados?*, se utiliza la *diversidad original* de cada consulta como margen, para luego estimar el porcentaje de palabras agregadas que llevan a una igual o mayor diversidad (Cuadros 5 y 6), utilizando el mismo número de snippets (N=3).

Cuadro 5. Proporción de expansiones con mayor diversidad G1

Consulta	**Diversidad**	**Modelo A**	**Modelo B**	**Modelo C**
1	2.909	0.76	0.76	0.78
2	2.699	0.50	0.58	0.52
3	2.855	0.86	0.82	0.92
4*	**2.576**	0.82	0.82	0.78
5	2.957	0.32	0.24	0.24
6	2.857	0.62	0.68	0.56
7	2.863	0.86	0.92	0.82
8	2.947	0.30	0.32	0.40
9	2.854	0.70	0.80	0.76
10	2.928	0.16	0.08	0.12
11	2.917	0.36	0.50	0.56
12	2.772	0.36	0.54	0.48
Resumen	2.845 (0.112)	0.551 (0.248)	**0.588 (0.288)**	0.578 (0.244)

Cuadro 6. Proporción de expansiones con mayor diversidad G2

Consulta	**Diversidad**	**Modelo A**	**Modelo B**	**Modelo C**
1	2.908	0.52	0.60	0.58
2	2.745	0.50	0.54	0.50
3	2.958	0.62	0.62	0.56
4	2.883	0.22	0.24	0.26
5*	**2.969**	0.32	0.34	0.32
6	2.910	0.76	0.76	0.62
7	2.961	0.28	0.22	0.22
Resumen	2.905 (0.077)	0.457 (0.190)	**0.474 (0.208)**	0.437 (0.165)

En general, ambos grupos tienen en promedio una alta diversidad. Para el Grupo 1 se tiene un valor promedio de diversidad de 2.845 (DE = 0.112) y para el Grupo 2, un valor de 2.905 (DE = 0.077).

Considerando el comportamiento en los extremos previo a la expansión, se tiene la mayor diversidad (+D) en la consulta 5 del Grupo 2 *"construir un auto"* (2.969), y la menor diversidad (-D) en la consulta 4 del Grupo 1 *"construir automóvil derby soapbox"* (2.576). Las Figuras 4 y 5 muestran la distribución de diversidad posterior de estos dos casos.

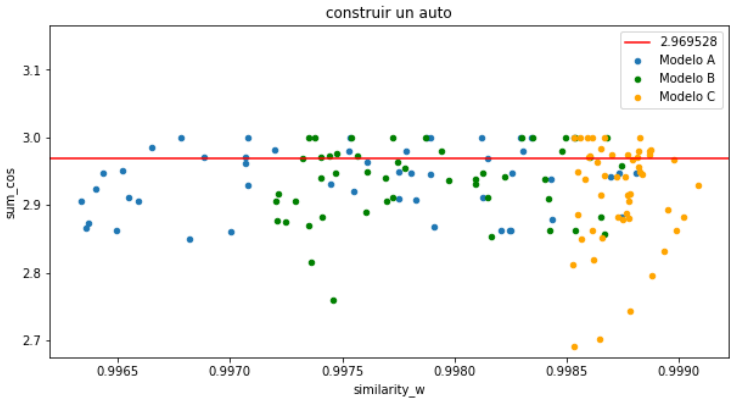

Figura 4. Diversidad en torno al máximo +D (Consulta 5 G2)

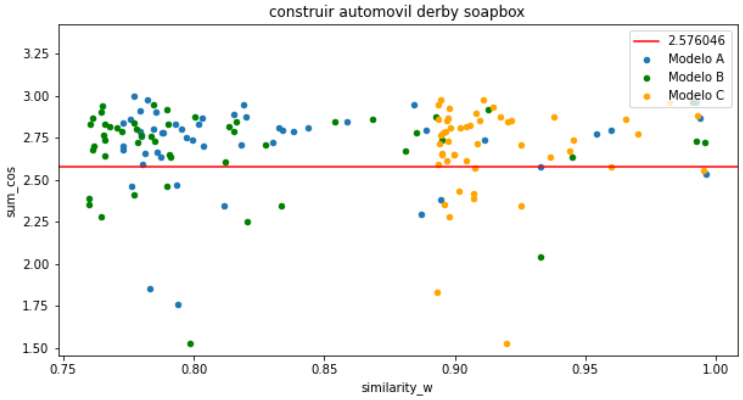

Figura 5. Diversidad en torno al mínimo -D (Consulta 4 G1)

En el extremo superior (+D) el modelo B contiene el mayor número de palabras sugeridas (34 %), para el extremo inferior (-D), el modelo A y B contienen el mismo porcentaje de palabras (82 %). El cuadro 7 contiene los 15 términos con mayor diversidad obtenida para los casos -D y + D.

Cuadro 7. Top 15 términos con mayor diversidad en los extremos

	+D			**-D**		
Consulta	"crear un auto básico"			"construir automóvil derby soapbox"		
Términos (t_i)	*debe*	*proceso*	*pueden*	*bull*	*vez*	*locos*
	básico	*realizar*	*elementos*	*noviembre*	*world*	*coches*
	principales	*tiempo*	*dentro*	*super*	*carros*	*annual*
	crear	*chasis*	*necesarios*	*box*	*racer*	*realiza*
	aprender	*plano*	*posible*	*blue*	*grid*	*santiago*

Globalmente el mejor modelo es el Modelo B, teniendo en promedio un 58.8 % y un 47.4 % de palabras que llevan a una diversidad mayor o igual a cada consulta original.

Previo a la expansión (Figura 3) se puede observar un comportamiento lineal entre la diversidad y el número de términos. Con respecto al modelo B se estudia la proporción promedio de expansiones en base al número de términos (Cuadro 8).

Cuadro 8. Proporción promedio en torno al número de términos

Términos (Consultas)	**Diversidad promedio**	**Proporción promedio**
1 (1)	2.700	0.58
2 (3)	**2.815**	**0.38**
3 (3)	**2.938**	**0.26**
4 (6)	2.839	0.67
5 (2)	2.883	0.78
6 (4)	**2.930**	0.50

Se tiene un mayor número de casos con una proporción igual o mayor al 50 % de palabras. Las consultas que se componían antes de 2 y 3 términos son la excepción y corresponden a casos sobre y bajo el promedio de diversidad original global 2.867 (DE = 0.100).

Para responder a la segunda pregunta de investigación *"¿Cómo se relacionan la similitud semántica y la diversidad estimada?*, se estudia el promedio de similitud *similarity_w* de las 50 palabras sugeridas por cada modelo (Cuadros 9 y 10).

Cuadro 9. Similitud promedio con el último término de la consulta G1

Consulta	**Modelo A**	**Modelo B**	**Modelo C**
1	0.998 (2.13×10^{-4})	0.997 (2.77×10^{-4})	0.998 (9.20×10^{-5})
2	0.828 (6.42×10^{-2})	0.816 (6.64×10^{-2})	0.915 (2.72×10^{-2})
3	0.993 (1.46×10^{-3})	0.995 (6.36×10^{-4})	0.998 (2.25×10^{-4})
4	0.828 (6.42×10^{-2})	0.816 (6.64×10^{-2})	0.915 (2.72×10^{-2})
5	0.997 (7.61×10^{-4})	0.997 (4.90×10^{-4})	0.998 (1.41×10^{-4})
6	0.998 (2.13×10^{-4})	0.997 (2.77×10^{-4})	0.998 (9.20×10^{-5})
7	0.996 (1.35×10^{-3})	0.997 (9.64×10^{-4})	0.998 (5.00×10^{-4})
8	0.997 (7.61×10^{-4})	0.997 (4.90×10^{-4})	0.998 (1.41×10^{-4})
9	0.997 (7.61×10^{-4})	0.997 (4.90×10^{-4})	0.998 (1.41×10^{-4})
10	0.993 (1.46×10^{-3})	0.995 (6.36×10^{-4})	0.998 (2.25×10^{-4})
11	0.998 (2.13×10^{-4})	0.997 (2.77×10^{-4})	0.998 (9.20×10^{-5})
12	0.828 (6.42×10^{-2})	0.816 (6.64×10^{-2})	0.915 (2.72×10^{-2})
Resumen	0.954 (1.66×10^{-2})	0.952 (1.69×10^{-2})	**0.977** (6.94×10^{-3})

Cuadro 10. Similitud promedio con el último término de la consulta G2

Consulta	**Modelo A**	**Modelo B**	**Modelo C**
1	0.997 (7.61×10^{-4})	0.997 (4.90×10^{-4})	0.998 (1.41×10^{-4})
2	0.995 (2.70×10^{-5})	0.999 (3.60×10^{-5})	0.994 (3.50×10^{-5})
3	0.993 (1.46×10^{-3})	0.995 (6.36×10^{-4})	0.998 (2.25×10^{-4})
4	0.933 (1.67×10^{-2})	0.948 (1.26×10^{-2})	0.980 (4.07×10^{-3})
5	0.997 (7.61×10^{-4})	0.997 (4.90×10^{-4})	0.998 (1.41×10^{-4})
6	0.998 (2.13×10^{-4})	0.997 (2.77×10^{-4})	0.998 (9.20×10^{-5})
7	0.995 (1.29×10^{-3})	0.994 (1.60×10^{-3})	0.997 (3.23×10^{-4})
Resumen	0.987 (3.04×10^{-3})	0.990 (2.31×10^{-3})	**0.996** (7.19×10^{-4})

Para ambos grupos la variación de similitud es relativamente pequeña, para G1 se tiene un promedio de 0.961 (DE = 0.011) y para G2 un promedio de 0.991 (DE = 0.003). En ambos casos se obtiene el valor más alto en el modelo C (0.977 y 0.996 respectivamente). Esto sugiere una relación directa entre la frecuencia de los términos permitida *min_count* y la similitud codificada por el modelo *similarity_w*.

Aunque es esperable que una reducción en el vocabulario del modelo cause un aumento en la similitud, una mayor similitud promedio no conlleva necesariamente a una mejor diversidad. Esto puede observarse en el caso del Modelo C, en G1 es el segundo modelo con mejor porcentaje de palabras (57 %) y para G2 es el último con un 43.7 %.

De forma similar al caso de los extremos, se estudia la distribución de diversidad sobre el modelo B (Figura 6) de todas las expansiones.

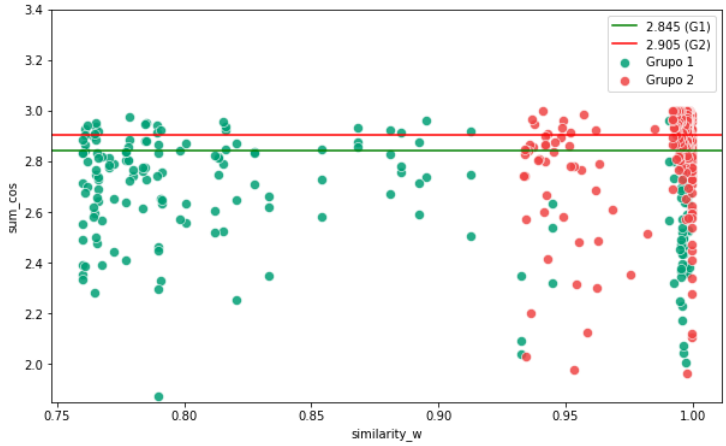

Figura 6. Diversidad posterior y similitud semántica del modelo B

Visualmente se puede apreciar una mayor desviación en torno a la similitud semántica y a la diversidad en el Grupo 1. Con respecto a la diversidad posterior, para ambos grupos el promedio es menor sobre la diversidad original 2.826 (DE = 0.214) y 2.856 (DE = 0.196) respectivamente. El Cuadro 11 describe con mayor detalle los resultados del Modelo B.

Cuadro 11. Resumen Modelo B

Descripción	**G1**	**G2**
Intervalo de similitud	[0.759 - 0.998]	[0.933 - 0.999]
Promedio de similitud	0.952 (1.69×10^{-2})	0.990 (2.31×10^{-3})
Intervalo de diversidad	[1.357 - 3]	[1.654 - 3]
Promedio de diversidad original	2.854 (0.112)	2.905 (0.077)
Promedio de diversidad posterior	**2.826 (0.214)**	**2.856 (0.196)**

En adición a esto, se quiere determinar dónde se encuentran los términos que ofrecen una mejor diversidad promedio. Para ello se agrupan en bloques de 10 todas las palabras sugeridas según su posición entre 0 y 50 (Figura 7). Tomando en cuenta que las palabras están en orden de similitud descendente, se obtienen los mejores promedios de diversidad a partir del término 20 (G1: 2.843, 2.855 ; G2: 2.859, 2.872 y 2.894).

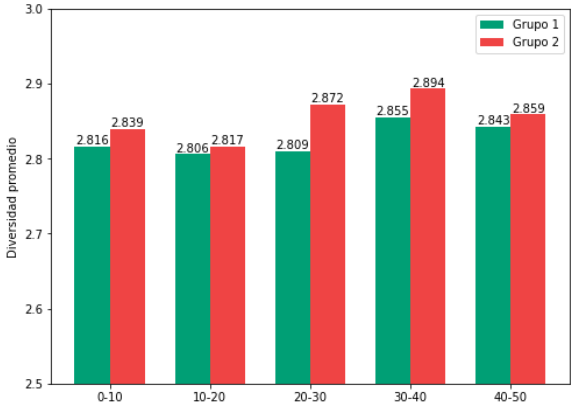

Figura 7. Diversidad promedio obtenida según el rango de similitud (Modelo B)

5. Discusión

Los resultados presentados en la sección anterior muestran la posibilidad de un aumento en la diversidad estimada *sum_cos* con la incorporación de un solo término semánticamente relevante.

Considerando 50 candidatos para cada par grupo-modelo se tiene entre un 43.7 % (21.85) y un 58.8 % (29.4) de casos con una mayor diversidad a la original.

Anterior a la expansión, se distingue una relación directa entre el número de términos y la diversidad obtenida (Figura 3), existiendo un punto de inflexión en las consultas de 3 términos.

Por un lado, la respuesta del buscador tiene sentido con la literatura, donde el número de términos y ambigüedad de una consulta, impacta en la diversidad de los resultados [15]. En el rango de 3 snippets, para una consulta de un solo término, se tiene en comparación una menor diversidad promedio (2.700) con el resto de consultas entre 2 y 6 términos (MIN= 2.815, MAX = 2.938).

Individualmente, las consultas del Grupo 1 tienen una menor diversidad promedio (2.845) que las del Grupo 2 (2.905). Esta diferencia puede interpretarse como el resultado de la interacción de los usuarios con el sistema IR, recordando que el Grupo 2 contiene las consultas más emitidas por los usuarios, sin considerar la primera consulta (Grupo 1).

Con respecto al promedio de términos, un 43 % de las consultas del Grupo 2 tienen originalmente 4 o más términos, para el Grupo 1 corresponde a un 75 %. Esto sugiere que la tendencia en la reformulación fue hacia la generalización de las consultas [3], en otras palabras, los usuarios removieron palabras para ampliar la búsqueda causando una mejora en la diversidad promedio.

Esto refleja la complejidad de la incerteza de las consultas, y el impacto que puede tener una sola palabra, pudiendo mejorar o empeorar la diversidad.

Esta poca estabilidad puede deberse también a la definición de la métrica. Teniendo cerca de 25 palabras en promedio por snippet (Cuadro 3), puede no ser suficiente una comparación léxica para definir diversidad, pese a la poca variación de palabras únicas que muestra el corpus (Cuadro 2). También podría extenderse el rango de snippets para precisar la diversidad de una consulta sobre otra.

En esencia, un valor bajo de diversidad significa que se están utilizando las mismas palabras para resumir los documentos en snippets. Si bien es posible mejorar la diversidad, es necesario estudiar en qué casos se logra y cómo se relacionan con la similitud semántica.

En los extremos, para el caso de mayor diversidad +D, se tiene la consulta *"construir un auto"* (G2) con un valor de diversidad de 2.969, y en el de menor diversidad -D *"construir automóvil derby soapbox"* (G1) con una diversidad de 2.576.

Sí se observa como una transición de G2 a G1, el caso -D puede ser una expansión de +D incorporando los términos *"derby"* y *"soapbox"*. Al precisar el contexto de la tarea de búsqueda, la similitud léxica de los primeros tres resultados aumenta en un 131 % .

Esto refuerza la idea de incorporar mecanismos para mejorar la estabilidad de la métrica, como por ejemplo, niveles de tolerancia que permitan asegurar la relevancia de los snippets con la tarea de búsqueda o utilizar la retroalimentación del usuario.

Luego de la expansión de ambos grupos, se evalúa qué modelo es mejor, en función del porcentaje de palabras agregadas que lleva a una misma o mejor diversidad. El mejor modelo es el B, conteniendo un 58.8 % y un 47.4 % de palabras con un valor de diversidad igual o mayor a cada consulta original.

En relación al número de términos (Cuadro 8), sólo las consultas originalmente de 2 y 3 términos tienen una proporción promedio menor al 50 % sobre el modelo B. Tomando en cuenta cada consulta de forma individual, solo en 6 de estas no se supera el 50 %. Para la totalidad de estos casos, la diversidad original es menor al promedio global de diversidad 2.867 (DE = 0.100).

Esto muestra que independiente del número de términos, cuando se tiene una diversidad bajo el promedio, es menor el número de palabras que permiten mejorarla.

En cuanto a la similitud semántica, la variación entre los modelos para cada grupo es baja. Para el Grupo 1 se tiene un promedio de 0.961 (DE = 0.011) y para el Grupo 2, se tiene un promedio de 0.991 (DE = 0.003)

Esto se puede atribuir partiendo desde la fuente de información, a la calidad del buscador y como este resume cada documento en un snippet, no dejando de lado las consideraciones establecidas de esta propuesta.

En primer lugar, sólo se está considerando una parte de la consulta para seleccionar los términos más próximos. En conjunto, los modelos sugieren sobre un 50 % de palabras únicas. De 150 palabras sugeridas, para el Grupo 1 se tiene un 54.8 % de palabras únicas, y para el Grupo 2 un 57.3 %. Una mayor

especificación del contexto de la consulta puede producir cambios en el orden de las palabras sugeridas para cada modelo.

Por otro lado, originalmente el corpus contiene 15019 palabras únicas (Cuadro 3), posterior a la definición de los modelos, los vocabularios que se forman son respectivamente de 2416, 1416 y 662 palabras.

Es decir, solo se está utilizando entre un 4 % (4.4077) y un 16 % (16.086) de las palabras originales. Esta reducción si bien es intencional y guiada por el análisis del corpus, podría estar descartando palabras relevantes que se repiten pero en diferentes conjugaciones y formas, debido a que no se considera en ningún grado la normalización de texto como una etapa en el preprocesamiento.

También el idioma es un factor a considerar en el preprocesamiento del corpus y en la selección de los términos, sobretodo cuando hay una predisposición sobre el tema de búsqueda al uso de palabras en inglés. Un claro ejemplo de esto ocurre en el caso de menor diversidad -D (Cuadro 7), donde el último término es *"soapbox"* y una parte de las mejores palabras son términos en inglés.

Los resultados permiten concluir que globalmente las consultas reformuladas tienen una mayor similitud semántica promedio y una mejor diversidad original.

Luego de la expansión, considerando el mejor modelo de diversidad se estudia el comportamiento entre la similitud semántica y diversidad posterior (Figura 6).

A través de la figura se aprecia de mejor forma la desviación mayor en similitud y diversidad del Grupo 1 sobre el Grupo 2. Como se mencionó anteriormente, una mayor similitud semántica no implica necesariamente una mejor diversidad, además cuando la diversidad es baja es menor el número de palabras que permiten mejorarla.

Por este motivo, se quiere identificar donde ocurre la mayor diversidad promedio en cada rango de similitud (Cuadro 11). Analizando independientemente los grupos, ocurre una disminución en la diversidad cercano al término 20 en adelante (Figura 7), incluso para el Grupo 2 que tenía una diversidad original menor que el Grupo 1.

Esto indica que aun con una baja diferencia de semántica, los términos con una menor similitud semántica son más propensos a llevar a resultados con mayor diversidad.

6. Conclusión

En este artículo se explora si la expansión de consultas basada en semánticas puede mejorar la diversidad de los resultados.

Esta hipótesis está acotada dentro de GoNSA2, una plataforma de aprendizaje basado en búsquedas, donde se tiene acceso a los registros de consulta e información contextual de las tareas de búsqueda.

En particular se estudia la tarea de búsqueda "Como construir un auto soapbox", expandiendo dos grupos de consultas mediante relevancia ciega como origen del corpus, junto al uso de 3 modelos Word2vec con la arquitectura Skip-Gram para la obtención de términos semánticamente relevantes.

Al analizar los resultados, se obtiene que el mejor modelo es el B conteniendo un 58.8 % y un 47.4 % de palabras que llevan a una igual o mejor diversidad léxica. Estos datos indican la posibilidad de una mejora en la diversidad con solo un término agregado. Además, en la selección de términos se muestra que una menor similitud semántica puede conducir a una mayor diversidad en los primeros resultados.

Cabe destacar que esta investigación representa un primer acercamiento a una implementación concreta sobre la plataforma GoNSA2, con la que se pueda dar soporte a la diversidad en la elaboración de consultas.

No obstante, para esto aún es necesaria una investigación mayor del modelo Word2vec y sus demás parámetros, junto a las etapas de preprocesamiento y medición de diversidad del método que se ven afectadas por una comparación exclusivamente léxica.

Referencias

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering **17**(6), 734–749 (2005)
2. Besbes, G., Baazaoui-Zghal, H.: Fuzzy ontologies for search results diversification: Application to medical data. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing. pp. 1968–1975 (2018)
3. Bilal, D., Gwizdka, J.: Children's query types and reformulations in google search. Information Processing & Management **54**(6), 1022–1041 (2018)
4. Bouchoucha, A., He, J., Nie, J.Y.: Diversified query expansion using conceptnet. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 1861–1864 (2013)
5. Câmara, A., Santos, R.L.: Traversing semantically annotated queries for task-oriented query recommendation. In: Proceedings of the 13th ACM Conference on Recommender Systems. pp. 511–515 (2019)
6. Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 243–250 (2008)
7. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 335–336 (1998)
8. Diaz, F., Mitra, B., Craswell, N.: Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605.07891 (2016)
9. Duarte Torres, S., Hiemstra, D., Serdyukov, P.: An analysis of queries intended to search information for children. In: Proceedings of the third symposium on Information interaction in context. pp. 235–244 (2010)
10. Duarte Torres, S., Hiemstra, D., Weber, I., Serdyukov, P.: Query recommendation for children. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 2010–2014 (2012)
11. Fails, J.A., Pera, M.S., Anuyah, O., Kennington, C., Wright, K.L., Bigirimana, W.: Query formulation assistance for kids: What is available, when to help & what kids want. In: Proceedings of the 18th ACM International Conference on Interaction Design and Children. pp. 109–120 (2019)
12. Gomaa, W.H., Fahmy, A.A., et al.: A survey of text similarity approaches. International Journal of Computer Applications **68**(13), 13–18 (2013)
13. Hu, S., Dou, Z., Wang, X., Sakai, T., Wen, J.R.: Search result diversification based on hierarchical intents. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 63–72 (2015)
14. Huang, Z., Cautis, B., Cheng, R., Zheng, Y., Mamoulis, N., Yan, J.: Entity-based query recommendation for long-tail queries. ACM Transactions on Knowledge Discovery from Data (TKDD) **12**(6), 1–24 (2018)
15. Jiang, D., Leung, K.W.T., Yang, L., Ng, W.: Query suggestion with diversification and personalization. Knowledge-Based Systems **89**, 553–568 (2015)
16. Kuzi, S., Shtok, A., Kurland, O.: Query expansion using word embeddings. In: Proceedings of the 25th ACM international on conference on information and knowledge management. pp. 1929–1932 (2016)

17. Lin, Y.S., Jiang, J.Y., Lee, S.J.: A similarity measure for text classification and clustering. IEEE transactions on knowledge and data engineering **26**(7), 1575–1590 (2013)
18. Madrazo Azpiazu, I., Dragovic, N., Anuyah, O., Pera, M.S.: Looking for the movie seven or sven from the movie frozen? a multi-perspective strategy for recommending queries for children. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. pp. 92–101 (2018)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 (2013)
21. Mitsui, M., Shah, C.: Multi-word generative query recommendation using topic modeling. In: Proceedings of the 10th ACM Conference on Recommender Systems. pp. 27–30 (2016)
22. Olivares-Rodríguez, C., Guenaga, M., Garaizar, P.: Automatic assessment of creativity in heuristic problem-solving based on query diversity (2017)
23. Ooi, J., Ma, X., Qin, H., Liew, S.C.: A survey of query expansion, query suggestion and query refinement techniques. In: 2015 4th International Conference on Software Engineering and Computer Systems (ICSECS). pp. 112–117. IEEE (2015)
24. Rehurek, R.: Word2vec embeddings. `https://radimrehurek.com/gensim/models/word2vec.html`, accessed: 2021-07-20
25. Rehurek, R.: Word2vec embeddings. `https://tedboy.github.io/nlps/generated/generated/gensim.models.Word2Vec.most_similar.html`, accessed: 2021-07-20
26. Roy, D., Paul, D., Mitra, M., Garain, U.: Using word embeddings for automatic query expansion. arXiv preprint arXiv:1606.07608 (2016)
27. Shajalal, M., Aono, M., Azim, M.A.: Aspect-based query expansion for search results diversification. In: 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR). pp. 147–152. IEEE (2018)
28. Shao, T., Chen, H., Chen, W.: Query auto-completion based on word2vec semantic similarity. In: Journal of Physics: Conference Series. vol. 1004, p. 012018. IOP Publishing (2018)
29. Tahery, S., Farzi, S.: Customized query auto-completion and suggestion—a review. Information Systems **87**, 101415 (2020)
30. Tamine-Lechani, L., Boughanem, M., Daoud, M.: Evaluation of contextual information retrieval effectiveness: overview of issues and research. Knowledge and Information Systems **24**(1), 1–34 (2010)
31. Umemoto, K., Yamamoto, T., Tanaka, K.: Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 405–414 (2016)
32. White, R.W., Richardson, M., Yih, W.t.: Questions vs. queries in informational search tasks. In: Proceedings of the 24th international conference on World Wide Web. pp. 135–136 (2015)