Anonimización de registros de búsqueda mediante la semántica de las consultas

Arnau Erola, Jordi Castellà-Roca
Departament d'Enginyeria Informàtica i Matemàtiques,
UNESCO Chair in Data Privacy,
Universitat Rovira i Virgili,
Av. Països Catalans 26, E-43007 Tarragona, Spain
Email: arnau.erola,jordi.castella@urv.cat

***Resumen*—Los motores de búsqueda en Internet almacenan y agrupan las consultas de sus usuarios. Esta información les permite ofrecer servicios avanzados (corrección de los términos de las búsqueda, desambiguar términos, etc.). No obstante, las consultas pueden contener información sensible acerca de los usuarios y permitir su identificación. Por estos motivos los registros de los motores búsqueda deben ofrecer privacidad. Pero esta protección puede suponer una pérdida importante de información. Lo ideal sería que los métodos de protección proporcionaran privacidad y a su vez mantuvieran la utilidad de los datos. Con este fin los registros son sometidos a procesos de anonimización selectivos, uno de ellos es la microagregación semántica. En este trabajo presentamos un nuevo sistema de microagregación semántica que recupera el significado de las consultas efectuadas sin separar sus términos. El nuevo método mantiene la privacidad de los usuarios y la utilidad de los datos.**

I. Introducción

Los motores de búsqueda agregan las consultas que realiza cada usuario y las utilizan para proporcionar mejores servicios, por ejemplo desambiguar los términos de las consultas (hay términos que pueden tener más de un significado) o para corregir errores en los términos.

Las consultas que los usuarios realizan en los motores de búsqueda proporcionan una fuente de información muy útil para estudiar el comportamiento de los usuarios (perfil, relaciones, etc.) o el cambio de tendencias. Por eso, no es de extrañar que tanto investigadores como empresas de marketing sean codiciosos por acceder a esta información.

Sin embargo, los propietarios de estos registros (los motores de búsqueda) son reticentes de proporcionarlos, debido a los problemas de privacidad que puede conllevar revelar esta información. Las consultas pueden contener datos sensibles como enfermedades, la tendencia sexual, las creencias religiosas, tecnologías en desarrollo (consultas realizadas desde el trabajo), la población de residencia, etc. Además, las consultas pueden identificar fácilemte a los usuarios si éstos se buscan a si mismos (consultas vanidosas), quedando su nombre en el registro [1].

De acuerdo con lo anterior, los registros de búsqueda no pueden ser publicados sin someterse previamente a un proceso de anonimización. La anonimización debe garantizar que las consultas (información) no pueden ser enlazadas con una identidad (usuario). Este proceso conlleva normalmente la eliminación de los identificadores (información que permite identificar a un usuario de forma unívoca), la información sensible y también la ofuscación de las consultas de los usuarios [2]. Pero estos métodos reducen la cantidad de información en los registros, es decir, reducen su utilidad [3].

No obstante, aunque los registros sean sometidos a un proceso de anonimización, si éste no se realiza correctamente, los registros de búsqueda anonimizados pueden contener suficiente información para identificar de manera unívoca algunos individuos. Un ejemplo de esto sucedió en 2006 [4], cuando la empresa AOL publicó unos 20 millones de consultas efectuadas por 658.000 usuarios durante 3 meses. El origen de la consulta (dirección IP) había sido sustituido por un pseudónimo y las consultas que contenían información sensible habían sido borradas. Aun así, un periodista del New York Times logró identificar a una mujer de Masachussets a patir de sus consultas.

I-A. Contribución y organización del artículo

En este trabajo proponemos un nuevo método de microagregación semántica. Al utilizar el significado completo de las consultas se consigue una mayor utilidad en los datos microagregados, manteniendo un elevado nivel de privacidad.

Se debe destacar que una consulta o un conjunto de consultas pueden revelar información sobre el usuario, pero éste puede considerar que esta información es privada o no según sus creencias. Por ejemplo, un usuario puede publicar su número de teléfono porque quiere que cualquier persona le pueda localizar (considera esta información como pública), mientras que otro usuario puede estar preocupado por recibir llamadas de gente que no conoce (considera esta información como privada). Determinar qué información es privada o no está fuera de la finalidad de este trabajo. En este trabajo, al igual que en trabajos del mismo ámbito [5], se considera que toda la información tiene la misma importancia y es privada.

En la Sección II se introduce un breve estado del arte. En la Sección III se describe brevemente la microagregación y el concepto de medida de semejanza usada en la microagregación. En la Sección IV se presenta la propuesta y en la

Sección V se muestran los resultados obtenidos. Finalmente la Sección VI contiene las conclusiones y el trabajo futuro.

II. Estado del arte

Una de las técnicas más utilizadas para anonimizar los registros de búsqueda consiste en la supresión de consultas [2]. Sin embargo, los métodos [4], [6], [7] basados en esta técnica no pueden garantizar que toda la información sensible haya sido eliminada, y generalmente, para prevenir la falta de privacidad, eliminan gran parte de la información de los registros, eliminando, a su vez, gran parte de su utilidad [8].

Por otro lado, la supresión de registros enteros implica que los motores de búsqueda no guarden información asociada a los usuarios, lo cual no es compatible con su sistema de negocio, ni tampoco ofrece ninguna utilidad a los investigadores.

Las técnicas de control de revelación estadística fueron usadas por primera vez en registros de búsqueda con medidas de semejanza basadas en la sintaxis de las consultas [9]. Los términos de las consultas son generalizados en [5] utilizando la herramienta WordNet [10]. En Erola et. al [3] los registros de consultas se microagregan semánticamente utilizando el Open Directory Project (ODP) [11]. Pero ODP solo permite clasificar términos, por lo que las consultas se dividen en términos y son éstos los que se clasifican en la estructura de ODP. Posteriormente, utilizando medidas de semejanza, se agregan los usuarios con mayor número de categorías parecidas.

III. Microagregación semántica

La microagregación [12] es un método de control de revelación estadística (Statistical Disclosure Control - SDC) que tiene como objetivo evitar que se pueda enlazar un registro con un usuario. Con ese fin, cada registro anonimizado comparte atributos con un mínimo de $k-1$ otros registros. Usualmente se consigue agrupando los registros de k usuarios y substituyendo cada registro por el centroide del grupo al que pertenece. La constante k es un parámetro del método que controla el nivel de privacidad. Cuanto mayor sea la k, mayor nivel de privacidad se obtiene, pero mayor es la pérdida de información.

Desde un punto de vista formal, la microagregación puede ser vista como un problema de agrupación (clustering), donde la utilidad se maximiza agrupando los registros más próximos [13]. Siendo los registros más próximos aquellos que tienen un mayor número de atributos en común. Así, los registros microagregados consiguen menores pérdidas de información.

Aunque generalmente estas técnicas son usadas con datos cuantitativos (numéricos), también pueden ser usadas para anonimizar los registros de búsqueda. En este caso es necesario definir una medida de semejanza entre consultas. Esta medida nos proporcionará el valor numérico necesario para efectuar la agrupación de los usuarios.

III-A. Semejanza en ODP

ODP [11] es el directorio más extenso de contenidos web editado por humanos. ODP está estructurado jerárquicamente en categorías y las páginas web están clasificadas en ellas. En la figura 1, se puede ver un ejemplo de esta clasificación. Las categorías están divididas en niveles, siendo los superiores los más genéricos.

Para medir la semejanza entre dos consultas clasificadas en ODP usaremos la medida de semejanza ODP_{sim}, propuesta en [3]. ODP_{sim} entre dos usuarios u_i y u_j se define de la siguiente manera:

$$OPD_{sim}(u_i,u_j)=\sum_{l=1}^{L}\{|c_l|:c_l\in\{C_l(u_i)\cap C_l(u_j)\}\} \quad (1)$$

Donde $C_l(u_i)$ es el conjunto de categorías para el usuario u_i en el nivel l y $|c_l|$ es el número de categorías para el usuario u_i en el nivel l, siendo $l\in\{1,\ldots,L\}$ la profundidad en la que trabajamos en el árbol ODP y L la profundidad máxima considerada.

IV. Microagregación de registros

La microagregación aplicada a los registros de consultas protege la privacidad de los usuarios. No obstante, la pérdida de información puede ser significativa en función de como se realice este proceso.

El método propuesto pretende crear grupos de k (o hasta $2k-1$) usuarios que compartan intereses en común, es decir, que los temas (categorías) de sus consultas sean lo más parecidos posibles. En este caso, el centroide (consultas asignadas al grupo de usuarios) que se obtendrá en la microagregación debería ser más parecido al registro de cada usuario.

Para agrupar a los usuarios que comparten más intereses primero debemos clasificar sus consultas. A partir de esta información podremos identificar a los usuarios que son más afines. Esta operación no es sencilla dado que cada consulta puede tener diversos términos, y el significado de un término puede ser diferente según los términos que le acompañan.

Para solventar este problema de ambiguación, utilizamos la ontología del própio motor de búsqueda, en nuestro caso Google, y después ODP. El motor de búsqueda proporciona la URL más significativa de la consulta, y ésta se clasifica en el árbol de ODP, obteniendo la categoría de la consulta. Al finalizar este proceso, para cada usuario se conocen sus categorías, de modo que se pueden agrupar los usuarios que comparten más categorías.

A continuación, se describen de forma detallada las fases que componen el método propuesto:

1. Obtención de las URLs asociadas a las consultas
2. Obtención de la categoría de la consulta
3. Agrupación de los usuarios
4. Obtención del centroide

En la descripción del método se utiliza la notación siguiente. Sea u_i un usuario, y n el número de usuarios de manera que U es el conjunto de usuarios, $U=\{u_1,\ldots,u_n\}$. El registro de

```
Open Directory Categories (1-5 of 100)
1. Recreation: Autos: Makes and Models: Audi (18)
2. Recreation: Autos: Makes and Models: Audi: Clubs (8)
3. Recreation: Autos: Makes and Models: Audi: A4 (5)
4. Recreation: Autos: Makes and Models: Audi: TT (4)
5. Recreation: Autos: Makes and Models: Audi: Clubs: United Kingdom (2)
```

Figura 1: Ejemplo de los resultados de una búsqueda en ODP.

consultas es $Q = \{Q_1, \ldots, Q_n\}$, donde $Q_i = \{q_{u_i}^1, \ldots, q_{u_i}^{p_i}\}$ son las consultas del usuario u_i.

IV-A. Obtención de las URLs asociadas a las consultas

Usar la ontología de Google para encontrar significados no es nuevo, ya fue usado en [14], [15]. Sin embargo, no tenemos constancia que haya sido usada anteriormente para anonimizar registros de búsqueda. Se ha seleccionado Google por su popularidad, pero se podría haber utilizado cualquier otro motor de búsqueda.

Esta parte del método es un preproceso de las consultas para la microagregación. Cada consulta del registro de búsqueda es enviada a Google. La primera URL diferente a un enlace promocionado o la wikipedia es almacenada. Se define $URL = \{URL(u_1), \ldots, URL(u_n)\}$ como el conjunto de resultados obtenidos en Google, siendo $URL(u_i) = \{url(_{u_i}^1), \ldots, url(_{u_i}^{p_i})\}$ el conjunto de enlaces del usuario u_i y $url(_{u_i}^j)$ los resultados seleccionados cuando se somete la consulta $q_{u_i}^j$.

ODP no puede tratar URLs completas. Por este motivo solo se almacenan los dominios de las URLs. Además, no todas las URLs dentro de un dominio están clasificadas en ODP. Los enlaces a Wikipedia no son almacenados porque contienen una gran variedad de temas y la URL $wikipedia.org$ siempre es clasificada en la misma categoría de ODP.

IV-B. Obtención de la categoría de la consulta

Una vez obtenidas las URLs debemos obtener sus categorías en ODP, es decir obtener $C_l(U) = \{C_l(u_1), \ldots, C_l(u_n)\}$. El algoritmo 1 muestra este proceso. A diferencia de [3], donde las consultas son separadas en términos y éstos son los que se clasifican en ODP, ahora clasificamos las URLs obtenidas haciendo una búsqueda inversa en el árbol ODP.

Una vez obtenida la clasificación en ODP se anonimizan los registros manteniendo la máxima homogeneidad en los datos. Con esa finalidad, se usa la expresión 1 que permite calcular la distancia entre consultas. Nótese que el valor máximo supone la máxima semejanza posible.

IV-C. Agrupación de los usuarios

A continuación se agrupan los usuarios más semejantes según su clasificación en ODP mediante el algoritmo 2.

IV-D. Obtención del centroide

Para cada grupo creado en la partición, se crea su cetroide a partir de consultas aleatorias de todos los usuarios que forman el grupo. El número de consultas con el que cada usuario contribuye al centroide depende del número de consultas que cada usuario del grupo tiene, y se puede expresar con la siguiente expresión:

$$Cuota_i = \frac{|Qi|}{k} \tag{2}$$

Donde $|Q_i|$ es el número de consultas del usuario u_i y k el tamaño del grupo.

Algoritmo 1 Algoritmo para clasificar las urls en ODP

Requiere: la máxima profundidad l en el árbol de ODP
Requiere: el conjunto de usuarios $U = \{u_i, \ldots, u_n\}$
Requiere: el conjunto de urls $URL(u_i) = \{url(_{u_i}^1), \ldots, url(_{u_i}^{p_i})\}$ para cada usuario u_i
Proporciona: el conjunto de categorias $C_l(u_i) = \{c_l(_{u_i}^1), \ldots, c_l(_{u_i}^{p_i})\}$ para cada usuario u_i
  para $u_i \in \{u_1, \ldots, u_n\}$ **hacer**
    para $url_{u_i}^j \in URLu_i = \{url(_{u_i}^1), \ldots, url(_{u_i}^{p_i})\}$ **hacer**
      obtener la categoria $c_l(_{u_i}^j)$ a profundidad l para $url(_{u_i}^j)$ usando ODP;
    fin para
  fin para
  devuelve C_l

Algoritmo 2 Algoritmo para agrupar a los usuarios

mientras el número de usuarios no agrupados sea mayor que $2k - 1$ **hacer**
  Se seleccionan los usuarios más semejantes (máximo ODP_{sim}) de entre todos los usuarios de U.
  mientras no se ha obtenido el tamaño k **hacer**
    Concatenamos los registros de los usuarios seleccionados.
    Buscamos el usuario que es más semejante con el registo concatenado.
  fin mientras
  Se crea un grupo con los usuarios seleccionados.
  Se borran los usuarios seleccionados de U.
fin mientras
Se crea un nuevo grupo con los usuarios restantes

V. Evaluación y resultados

En la evaluación se ha estudiado el nivel de privacidad y el nivel de utilidad de los registros una vez protegidos.

V-A. Medida de utilidad

Para evaluar los efectos de la microagregación semántica en la utilidad de los datos se ha utilizado la medida SRP propuesta en [3].

$$SRP = \frac{\rho}{\chi} \tag{3}$$

Donde, para una profundidad determinada l, ρ es el número de consultas de cada categoría que aparece en el registro original y a su vez en el centroide, y χ es el número original de consultas en l.

V-B. Resultados

A continuación se presentan los resultados obtenidos con el método propuesto y se comparan con el propuesto en [3].

V-B1. Privacidad: El resultado del proceso de microagregación es un registro que contiene las consultas de los centroides que se han creado a partir de las agrupaciones de k (o $2k-1$) usuarios. La microagregación proporciona k-anonimidad [16,17] a nivel de usuario, por lo que la privacidad del usuario es proporcionada por los $k-1$ usuarios restantes.

V-B2. Utilidad: Se han ejecutado los tests usando un conjunto de 840 usuarios seleccionados de forma aleatoria de los ficheros de AOL. Cada usuario realiza entre 400 y 600 consultas, por lo que tenemos cerca de $400,000$ consultas.

En la figura 2 se puede observar el SRP obtenido usando el método propuesto y el presentado en [3]. El nuevo método consigue cerca de un $10\,\%$ más de SRP, mejorando los valores de SRP en todas las profundidades y tamaños de los grupos de k usuarios. Los valores SRP son más altos en los niveles superficiales, ya que se pueden enlazar más consultas con categorias más generales, y menos en los niveles más profundos, ya que las categorías son más específicas y es más difícil de encontrar coincidencias.

En la figura 3 se puede observar el SRP obtenido por los 100 usuarios con mayor semejanza. En este caso el método propuesto logra una mejora del $21\,\%$ del SRP respecto a la microagregación presentada en [3]. El método propuesto podría obtener mejores resultados cuando los usuarios son más semejantes.

VI. Conclusiones

Los registros de búsqueda deben ser anonimizados para proteger la privacidad de los usuarios. Sin embargo, este proceso supone una pérdida de utilidad.

En este trabajo se ha propuesto un método de microagregación que utiliza la ontología de Google para agrupar a los usuarios con más intereses en común.

Los resultados muestran que el método reduce la perturbación introducida en los datos, proporcionando mayor utilidad que la microagregación propuesta en [3]. De este dato concluimos que las técnicas de microagregación de registros de búsqueda que descomponen las consultas en términos con el fin de buscar sus significados, como [3], no interpretan correctamente el significado de totas las consultas. Ésto produce una mayor pérdida de utilidad.

Además, se ha comprobado que el sistema puede ofrecer mejores resultados si los usuarios comparten más intereses. Éste podría ser el caso de los usuarios de una red social, donde los usuarios comparten muchos intereses en común con sus vecinos.

Una amenaza de interés particular en los registros de búsqueda microagregados es la desambiguación. Si las consultas contenidas en el centroide no tienen intereses en común, éstas pueden ser divididas, creando pequeños perfiles reales de usuarios. Como trabajo futuro estudiaremos como esta amenaza se comporta con los métodos existentes en la literatura y con el propuesto.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio español de Ciencia e Innovación [TSI2007-65406-C03-01, ARES CONSOLIDER INGENIO 2010 CSD2007-00004, Audit Transparency Voting Process IPT-430.000-2.010-31, CO-PRIVACYTIN2011-27076-C03-01], el Ministerio español de Industria, Comercio y Turismo [eVerification2 TSI-020100-2011-39, SeCloud TSI-020302 a 2010-153], y la Generalitat de Catalunya [2009 SGR1135]. Los autores son los únicos responsables de las opiniones expresadas en este documento, que no reflejan necesariamente la posición de la UNESCO ni comprometen la organización.

Referencias

[1] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. Vanity fair: privacy in querylog bundles. In *CIKM*, pages 853–862, 2008.

[2] A. Cooper. A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Transactions on the Web*, 2(4), 2008.

[3] Arnau Erola, Jordi Castell'a-Roca, Guillermo Navarro-Arribas, and Vicenc Torra. Semantic microaggregation for the anonymization of query logs using the open directory project. *SORT-Statistics and Operations Research Transactions*, 35(Special issue):25–40, Sep 2011.

[4] Eytan Adar. User 4XXXXX9: Anonymizing Query Logs. In *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007)*, 2007.

[5] Y. He and J. Naughton. Anonymization of set-valued data via top-down, local generalization. *Proceedings of the VLDB Endowment*, 2(1):934–945, 2009.

[6] B. Poblete, M. Spiliopoulou, and R. Baeza-Yates. Website privacy preservation for query log publishing. In *First International Workshop on Privacy, Security and Trust in KDD (PinKDD 2007)*, pages 80–96, 2008.

[7] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 171–180, 2009.

[8] Li Xiong and Eugene Agichtein. Towards Privacy-Preserving Query Log Publishing. In Einat Amitay, Craig G. Murray, and Jaime Teevan, editors, *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007)*, May 2007.

[9] G. Navarro-Arribas and V. Torra. Tree-based microaggregation for the anonymization of search logs. In *WI-IAT '09: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 155–158, 2009.

[10] G. Miller. WordNet - about us. WordNet. Princeton University, 2009.

[11] ODP. Open directory project, 2011.

[12] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada*, pages 195–204, 1993.

[13] J. Domingo-Ferrer and J.M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14:189–201, 2002.

[14] Rune Sætre, Amund Tveit, Tonje Strommen Steigedal, and Astrid Lægreid. Semantic annotation of biomedical literature using google. In *ICCSA (3)*, pages 327–337, 2005.

[15] Risto Gligorov, Zharko Aleksovski, Warner Kate, and Frank Van Harmelen B. Using google distance to weight approximate ontology matches. In *In: Proc. WWW-07. (2007*, pages 767–776. ACM Press, 2007.

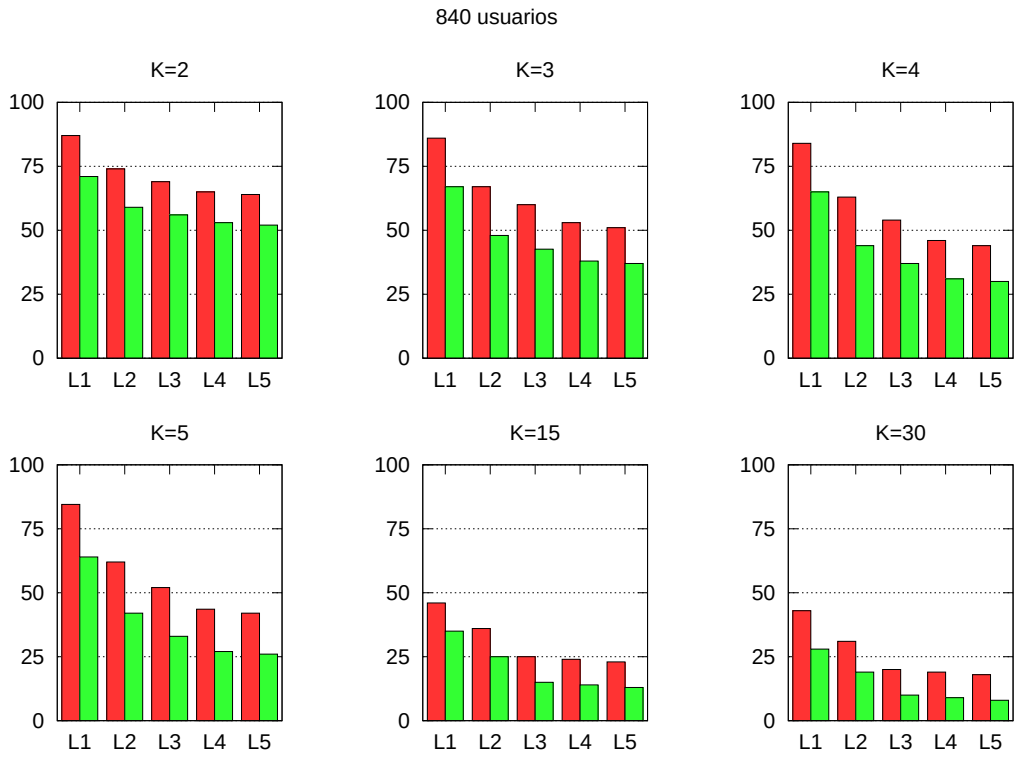

Figura 2: Porcentaje de semejanza semántica de los registros microagregados usando varias profundidades L en ODP.

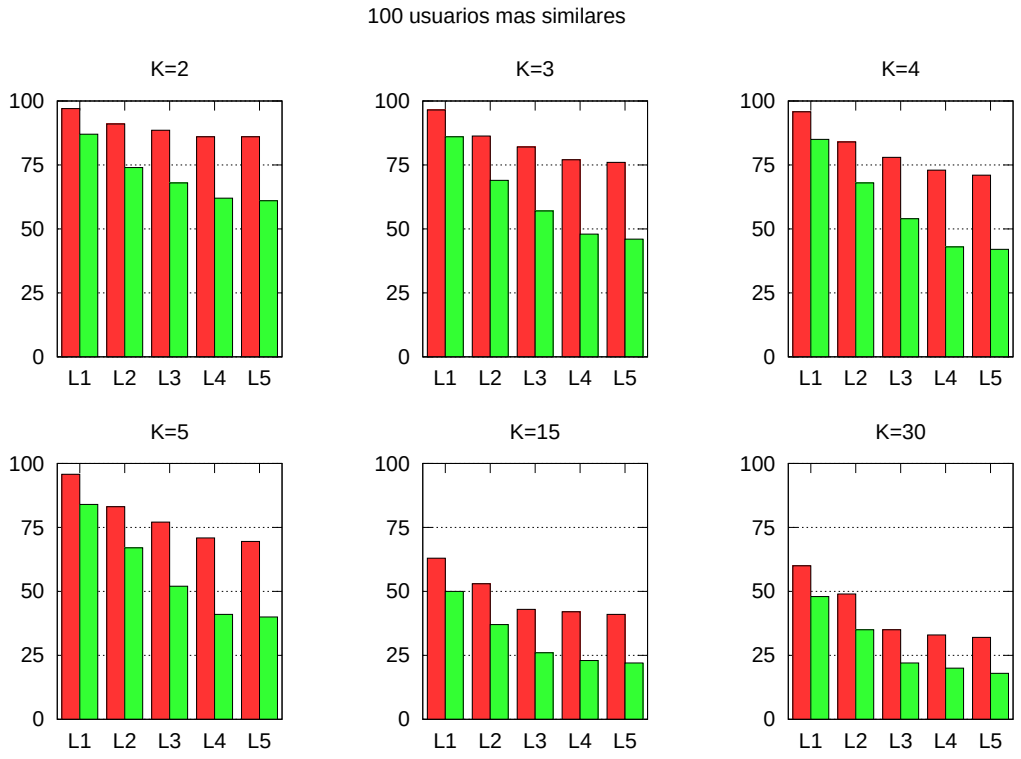

Figura 3: Porcentaje de semejanza semántica de los registros microagregados usando varias profundidades L en ODP.