Evaluación de la Calidad de la Información de Wikipedia en Español

Lian Pohn, Edgardo Ferretti, Marcelo Errecalde

Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Departamento de Informática, Universidad Nacional de San Luis
Ejército de los Andes 950 - (D5700HHW) San Luis - Argentina
e-mails:lian_pohn@hotmail.com,{ferretti,merreca}@unsl.edu.ar

Resumen

Este artículo describe, brevemente, las tareas de investigación y desarrollo que se están llevando a cabo para evaluar la calidad de información en la Web en el marco del proyecto "Herramientos y mecanismos para la toma de decisiones en agentes inteligentes artificiales". En particular, se ha tomado como caso de estudio primario la enciclopedia online Wikipedia en español. Este tema de trabajo permite la interacción de las dos líneas de investigación que contiene este proyecto y además se está abordando en forma conjunta con investigadores de Alemania, España y México, en el contexto de un proyecto FP7 financiado por la Unión Europea.

Palabras clave: Calidad de Información en la Web, Wikipedia, Sistemas Inteligentes.

Contexto

El tema motivo de esta presentación se encuentra dentro de los alcances de la línea de investigación "Minería de Textos y de la Web" del proyecto "Herramientos y mecanismos para la toma de decisiones en agentes inteligentes artificiales", Proyecto de Investigación consolidado de la Universidad Nacional de San Luis. En este proyecto, el objetivo principal es avanzar en la integración de las investigaciones sobre herramientas para la extracción y análisis inteligente de contenido Web de calidad y desarrollo e integración de modelos y mecanismos efectivos para la toma de decisiones, y el aprendizaje automático.

Este proyecto posee además otra línea de investigación: "Agentes Inteligentes"; y la articulación entre ambas líneas está dada por la meta de implementar un sistema inteligente automático que evalúe la calidad de la información en Wikipedia y que siga el paradigma de agentes.

Las dos fuentes de financiamiento de este proyecto son la Universidad Nacional de San Luis y la Comisión Europea de Investigación e Innovación, a través del programa Marie Curie Actions: FP7 People 2010 IRSES.

Introducción

En la actualidad, el acceso a la información es un tema clave en todos los aspectos relacionados con la vida moderna. En este sentido, la evaluación de la calidad de la información en la Web se ha convertido en una tarea crucial, dado que cada día son más las personas y entidades gubernamentales o privadas que toman decisiones basándose en información disponible en la Web. Asimismo, el notable incremento de información disponible en la Web ha potenciado la necesidad de eva-

luar su calidad de forma automática. Este hecho se debe entre otras cosas, a la creciente popularidad de sitios que permiten a usuarios comunes generar de forma muy sencilla contenido web y la inevitable divergencia en la calidad del contenido producido [8].

En este contexto, Wikipedia es un emprendimiento paradigmático. Esta enciclopedia de libre acceso generada a partir de las contribuciones de millones de usuarios, tiene esta característica como principal fortaleza en lo que respecta a su creciente popularidad, siendo (la versión inglesa) uno de los diez sitios más visitados en el mundo, en la actualidad. Sin embargo, esta característica es probablemente, el mayor desafío que Wikipedia enfrenta en cómo mejorar de forma sistemática la calidad informativa de sus archivos. Este aspecto no es casual si consideramos que los autores que contribuyen con Wikipedia son heterogéneos, en cuanto al nivel de educación, edad, cultura, habilidades del lenguaje y especialización en un área.

Una interpretación ampliamente aceptada de calidad de información, es que en sí mismo, es un concepto multi-dimensional que se define por ciertos aspectos de calidad (dimensiones); como por ejemplo: la exactitud, fiabilidad y relevancia [19]. Asimismo, la evaluación de la calidad de información requiere la consideración del contexto y casos de uso [23]. En particular, en el contexto de Wikipedia, es decir, el género de las enciclopedias, el ideal en lo que respecta a calidad de información ha sido formalizado bajo el concepto de artículo destacado (AD) (en inglés, *featured article*). Los ADs son artículos de Wikipedia que, como resultado de un arduo proceso de revisión entre pares, son considerados como artículos de alta calidad de acuerdo a criterios claramente definidos en la comunidad de Wikipedia. Estos criterios incluyen aspectos tales como la calidad de la escritura, la comprensibilidad, la existencia de un buen trabajo de investigación que lo sustente, neutralidad, estabilidad, entre otros.[1]

A pesar de lo importante que es contar con una definición precisa sobre qué constituye un AD, como se indica en [3], en la actualidad menos de 0,1 % de los artículos en `Wikipedia.org` son AD. En el caso de Wikipedia en español, se estima que este porcentaje es de 0,11 %.[2]

En la siguiente sección, se describen los principales enfoques desarrollados por la comunidad científica en lo que respecta al estudio de calidad de información en Wikipedia.

Líneas de Investigación y Desarrollo

Los métodos utilizados para determinar aspectos de calidad en los artículos de Wikipedia son muy variados; no obstante, en términos generales podrían identificarse tres líneas de investigación principales, que se describen a continuación.

Identificación de Artículos Destacados

Dado que los artículos de Wikipedia se escriben de forma colaborativa y se mantienen principalmente por voluntarios, la lógica detrás de la idea: “mientras mayor sea el número de ediciones de un artículo y mayor sea el número de editores, mayor debería ser su calidad”, es muy razonable. De hecho, Wilkinson y Huberman [24] proveen evidencia que los artículos destacados pueden ser distinguidos de aquellos que no lo son considerando el número de ediciones que han tenido y de editores distintos que han contribuido.

De la misma manera, continuando con la intuición anterior, mientras más ediciones tenga un artículo tentativamente éstos tenderán a tener más contenido textual. Este hallazgo también ha sido explorado por Blumenstock [9] que demostró experimentalmente que una característica tan sencilla como el número de palabras en el documento puede competir con características sofisticadas para

[1] `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`.

[2] `http://es.wikipedia.org/wiki/Wikipedia:Articulos_destacados`

distinguir artículos destacados de aquellos que no lo son.

El hecho de que los artículos destacados tienden a ser más largos hace que probablemente éstos tengan más contenido factual. Basándose en esta suposición, Lex et al. [15] propusieron una medida llamada *densidad factual* para discriminar entre artículos destacados y los que no lo son. Básicamente, esta medida relaciona el número total de hechos en un artículo con respecto a su longitud. Esta medida probó ser muy útil para discriminar artículos destacados cuya longitud era en promedio similar a la de los artículos comunes. Este estudio demostró que los artículos destacados tienden a ser más informativos que los que no lo son, independientemente de su longitud.

De acuerdo con los resultados presentados en [7, 16], este fenómeno se debe a que la producción conjunta y cooperación entre miembros de una comunidad Wiki logra que surja contenido de alta calidad.

Desarrollo de métricas y características para la evaluación de la calidad

Si bien los trabajos [7, 16], no tienen como objetivo particular la identificación de artículos destacados en Wikipedia, las conclusiones obtenidas de los mismos sirven de soporte para explicar los resultados obtenidos por Lex et. al. [15], como se comentó anteriormente.

Un trabajo análogo a [16] es el llevado a cabo por Hu et al. [14], donde se desarrollan modelos para medir la calidad de los artículos editados en Wikipedia basados en el principio: "buenos autores escriben buenos artículos y los buenos artículos están escritos por buenos autores", es decir, que existe una dependencia mutua entre calidad y autoría. Asimismo, en este trabajo también se reporta que a veces evaluar solamente la interacción entre editores ([7, 16]) no es suficiente para determinar la calidad de un artículo y que la longitud en los mismos tiende a aportar calidad informativa, sin que eso conlleve a que el mero hecho de que un artículo sea extenso implicará que sea de buena calidad.

En [20], Stvilia et al. proponen siete métricas computacionales específicas de calidad de información en Wikipedia. Probaron la validez de las mismas en la discriminación de artículos destacados de los que no lo son, con una exactitud cercana al 90 %. También, toman la iniciativa en la construcción de métricas de calidad de información que siendo originalmente pensadas para Wikipedia, se pueden adaptar para su aplicación en contextos similares de creación de contenido por parte de los usuarios.

Detección de Defectos de Calidad

A diferencia de las líneas de investigación descritas precedentemente, la detección de defectos de calidad (en inglés: *quality flaws*) intenta identificar imperfecciones o fallas de calidad específicas en los artículos de Wikipedia, convirtiéndose en los últimos tiempos en una de las tareas seleccionadas para su evaluación en la competencia PAN del CLEF 2012 Evaluation Labs and Workshop [2]. Si bien se pueden mencionar algunos trabajos que han seguido esta línea de investigación considerando pequeñas muestras de artículos [21] o analizaron solamente un conjunto restringido de deficiencias de calidad [5, 13], los primeros en realizar un análisis detallado de los defectos de calidad en Wikipedia fueron investigadores de la Universidad de Weimar encabezados por Anderka [1, 3, 6]. Este análisis revela que un 27,52 % de los artículos de Wikipedia en inglés contiene al menos un defecto de calidad y que el 70 % de las deficiencias se relacionan con la verificabilidad del artículo. Este análisis está basado en artículos etiquetados manualmente por lo que se supone que el número real de fallas es más alto. Por lo tanto, es altamente probable que muchos artículos con deficiencias no hayan sido identificados aún.

Respecto a la competencia llevada a cabo en el contexto del PAN [2], los organizadores plantearon la predicción de defectos de calidad en Wikipedia como un problema de clasificación de una clase [22] (*one-class classification problem*), como ya fuera propuesto

en [4, 6]. En este contexto, el problema es: dado un conjunto de artículos de Wikipedia etiquetados con un defecto de calidad particular, decidir si un artículo (no etiquetado) sufre de este defecto. En este problema, y como se resalta en [2], el desafío clave es la ausencia de datos de entrenamiento "negativos", es decir, artículos etiquetados como "no conteniendo" una falla particular. Esto hace que ciertas técnicas clásicas de clasificación basadas en discriminación (binaria o multiclase) sean inaplicables. Por lo tanto, la ingeniería de características (*feature engineering*), es decir, el desarrollo de modelos de documentos que discriminan artículos conteniendo una cierta falla desde todos los otros artículos, se convierte en un factor crucial.

De hecho, el equipo ganador [11] de esta competencia implementó un sub-conjunto de las features propuestas en [6] para la representación de los artículos de Wikipedia y enfatizó el estudio experimental con distintas variantes algorítmicas de PU Learning [18] relacionadas principalmente con la evaluación de distintas estrategias de muestreo de documentos no etiquetados y diferentes enfoques para la selección de negativos confiables.

Resultados y Objetivos

De acuerdo con lo expuesto anteriormente, puede observarse que si bien existe un área de investigación muy activa relacionada a la calidad de la información en Wikipedia, estos estudios están centrados casi mayoritariamente en idioma inglés, no existiendo de acuerdo a nuestro conocimiento, estudios similares con Wikipedia en español.

Los antecedentes de nuestro grupo en esta temática [11, 15] se han basado en la versión inglesa de Wikipedia. Es por eso, que se tiene como objetivo general realizar una primera aproximación al problema del análisis de la calidad de la información de Wikipedia en español. En particular, nos centraremos en aspectos vinculados al soporte para la categorización automática de ADs y de algunos defectos de calidad más frecuentes.

En este contexto, un objetivo parcial a cumplir será el estudio y análisis comparativo del estado de Wikipedia en español con respecto a otra más desarrollada como es Wikipedia en inglés. También se estudiará la factibilidad de crear un pequeño corpus de entrenamiento con artículos destacados para verificar si alguna de las conclusiones obtenidas con técnicas sencillas como las utilizadas en [9] y [17] también se verifican con los artículos de Wikipedia en español. Finalmente, como un objetivo secundario, se prevé la realización de un estudio preliminar tendiente a analizar y comparar la situación de los defectos de calidad de Wikipedia en español con respecto a su versión en inglés y, en caso de ser posible realizar un trabajo inicial con alguno de los defectos más frecuentes.

Formación de Recursos Humanos

Trabajos de tesis vinculados con las temáticas descritas previamente:

- 1 tesis de Licenciatura en ejecución.

Referencias

[1] M. Anderka and B. Stein. A Breakdown of Quality Flaws in Wikipedia. In Castillo et al. [10], pages 11–18.

[2] M. Anderka and B. Stein. Overview of the 1th international competition on quality flaw prediction in wikipedia. In Forner et al. [12].

[3] M. Anderka, B. Stein, and M. Busse. On the Evolution of Quality Flaws and the Effectiveness of Cleanup Tags in the English Wikipedia. In *Wikipedia Academy 2012*. Wikipedia, July 2012.

[4] M. Anderka, B. Stein, and N. Lipka. In *20th ACM International Conference on Information and Knowledge Management (CIKM 2011)*. ACM.

[5] M. Anderka, B. Stein, and N. Lipka. Towards automatic quality assurance in wikipedia. In *20th International Conference on World Wide Web*, 2011.

[6] M. Anderka, B. Stein, and N. Lipka. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In *35th Annual SIGIR Conference*. ACM, 2012.

[7] D. Anthony, S. Smith, and T. Williamson. Reputation and reliability in collective goods: The case of the online encyclopedia wikipedia. *Rationality & Society*, 21(3), 2009.

[8] R. Baeza-Yates. User generated content: how good is it? In *3rd Workshop on information credibility on the Web*, 2009.

[9] J. E. Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *17th International Conference on World Wide Web*, 2008.

[10] C. Castillo, Z. Gyongyi, A. Jatowt, and K. Tanaka, editors. *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*. ACM, 2012.

[11] E. Ferretti, D. H. Fusilier, R. Guzmán-Cabrera, M. M. y Gómez, M. Errecalde, and P. Rosso. On the use of pu learning for quality flaw prediction in wikipedia. In Forner et al. [12].

[12] P. Forner, J. Karlgren, and C. Womser-Hacker, editors. *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, 2012.

[13] L. Gaio, M. den Besten, A. Rossi, and J. Dalle. Wikibugs: using template messages in open content collections. In *5th Symposium on wikis and open collaboration*, 2009.

[14] M. Hu, E. Lim, A. Sun, H. Lauw, and B. Vuong. Measuring article quality in wikipedia: models and evaluation. In *CIKM*. ACM, 2007.

[15] E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein, and M. Granitzer. Measuring the quality of web content using factual information. In Castillo et al. [10].

[16] A. Lih. Wikipedia as participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource. In *5th international symposium on online journalism*, 2004.

[17] N. Lipka and B. Stein. Identifying Featured Articles in Wikipedia: Writing Style Matters. In *19th International Conference on World Wide Web*, 2010.

[18] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *3rd IEEE International Conference on Data Mining*, 2003.

[19] T. Redman. *Data Quality for the Information Age*. Artech House, 1996.

[20] B. Stvilia, M. Twidale, L. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *ICIQ*, 2005.

[21] B. Stvilia, M. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in wikipedia. *JASIST*, 59(6):983–1001, 2008.

[22] D. Tax. *One-class classification*. PhD thesis, Technische Universiteit Delft, 2001.

[23] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 1996.

[24] D. Wilkinson and B. Huberman. Cooperation and quality in wikipedia. In *3rd International symposium on wikis and open collaboration*. ACM, 2007.