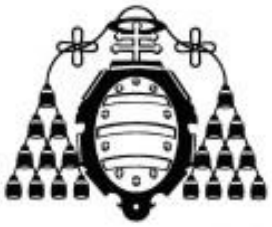
UNIVERSIDAD DE OVIEDO



ESCUELA DE INGENIERÍA INFORMÁTICA

PROYECTO FIN DE MÁSTER

"ACOTA: TECNOLOGÍAS DE ETIQUETADO SEMIAUTOMÁTICO Y COLABORATIVO"

Vº Bº del Director del Proyecto	**AUTOR: César Luis Alvargonzález** **DIRECTOR: Jose Emilio Labra Gayo**

Dedicado a mi abuela Montse y a mi abuelo alejandro por todo el apoyo que me dieron, sin los que no habría podido llegar a donde estoy.

Agradecimientos

Agradecer a mis amigos y familiares por estos últimos meses de ausencia centrado en el desarrollo de este proyecto, ya que siempre han estado encima de mí apoyándome y reconfortándome siempre que ha sido necesario.

Quiero agradecer a mi director por estar presente y por brindarme la oportunidad de formar parte del grupo de investigación WESO, sin el que este proyecto no habría sido desarrollado. También no quiero olvidarme de Jose María Álvarez-Rodríguez que de forma desinteresada, estuvo siempre pendiente del desarrollo del proyecto como de las diferentes publicaciones. También quiero agradecer al resto de integrantes del grupo de investigación WESO, especialmente a Weena cuyo trabajo previo me sirvió como pilar sobre el que asentar este trabajo de investigación.

Quiero a gradecer también a la empresa Treelogic S.L. por servirnos el conjunto de datos que sirvieron como muestra sobre los que se realizaron la evaluación de este trabajo.

Además agradecer a FEDER/CDTI por brindar el marco económico sobre el que pudo desarrollarse este proyecto, formando parte del proyecto ORIGIN (ORganizaciones Inteligentes Globales INnovadoras), cofinanciado por FEDER/CDTI (Fondos europeos al desarrollo regional) con el objetivo de aumentar la productividad de las actividades de desarrollo de software en escenarios globales.

Resumen

A pesar de que ha existido un gran número de trabajaos enfocados en el desarrollo de técnicas de recomendado automático y/o social, dichos componentes suelen estar enfocados en idiomas en concreto (generalmente en inglés), existiendo poca investigación centrada en técnicas de este tipo que permitan procesar contenidos multilingües.

Este trabajo presenta una metodología multilingüe hibrida semiautomática y colaborativa que combina técnicas de etiquetado automático con técnicas de recomendación de etiquetas basadas en el comportamiento previo de los usuarios con el sistema. Además se ha presenta una implementación de referencia llamada ACOTA (Automatic Collaborative Tagging) con el fin de demostrar las funcionalidades de recomendación aportadas que permiten asistir a usuarios, tanto nóveles como expertos, a la hora de etiquetar recursos multilingües. Por último, se ha desarrollado un estudio en el contexto de gestión del conocimiento empresarial, con el fin de evaluar la precisión y calidad del funcionamiento de la metodología propuesta.

Palabras Clave

Etiquetado Colaborativo, Etiquetado Automático, Datos Multilingüe, Recuperación Documental, Gestión del Conocimiento

Abstract

Despite of automatic, social and collaborative tagging techniques have been the topic of certain number of recent studies, in the meantime, multilingual tagging techniques has been poorly noticed.

The present paper introduces a multilingual hybrid methodology to automatically deploy and combine collaborative tagging techniques based on user-behaviour and well-known recommendation algorithms to name a few. A reference implementation called ACOTA (Automatic Collaborative Tagging) is also outlined in order to demonstrate the novel recommendation capabilities of this approach with the aim of assisting both expert and non-expert users when multilingual resource tagging is required. Finally a quantitative research study in the context of corporate knowledge management is also presented to evaluate and assess the goodness and accuracy of the methodology to minimize the effort of multilingual document categorization

Keywords

Collaborative Tagging, Automatic Tagging, Multilingual Data, Document Retrieval, Knowledge Management

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

La gran cantidad de datos disponibles en Internet, así como el gran aumento del número de usuarios de dispositivos móviles como tablets, teléfonos inteligentes (smart-phones) y lectores de libros electrónicos, está propiciando un ecosistema en el que se genera una gran cantidad de contenidos digitales, en el que nuevos requisitos se están haciendo presentes [1]. Hace años dichos recursos de información era impresos en papel y almacenados en archivadores, pero en la actualidad dicha información [2] (en 2012 se producían 2.5 exabytes cada día en Internet) es almacenada en formatos digitales utilizando diferentes métodos de organización y técnicas de acceso, con el objetivo de facilitar tareas de extracción de información, de búsqueda y acceso, entre otras. A pesar de que se trata de hacer uso de las ventajas que facilitan el indexado y diferentes algoritmos de categorización, los recursos provenientes de Internet, suele ser un contenido altamente heterogéneo [3], incrementando la dificultad a la hora de procesar dicha digital suele ser multilingüe, es decir está en diferentes idiomas, siendo una de las principales hándicaps o retos a la hora de garantizar el acceso a dicha información [4]. En este sentido, el contenido disponible en Internet está conformado por una gran variedad de idiomas, aunque el inglés es el idioma más hablado [5], con un 54.9% de páginas web, estando la mayoría de técnicas y herramientas construidas en torno a dicho lenguaje. Esto implica que el resto de idiomas (45.1%) encara una falta de herramientas apropiadas para explotar este tipo de información. Teniendo en cuenta que hay alrededor de 14210 millones de páginas web indexadas [6], alrededor de 6440 millones de páginas web están en idiomas diferentes al inglés, por lo tanto existe una creciente necesidad de aportar nuevas funcionalidades y tecnologías para salvar la distancia existente de recursos de información ingleses y no ingleses.

Organizar esta ingente cantidad de información resulta una tarea compleja, repetitiva y por ende tediosa, pero en algunos ámbitos de aplicación, es una tarea crucial, como pueden ser los ambientes B2B (Bussines to Bussines), gestión del conocimiento empresarial, sistemas de extracción de información, sistemas de aprendizaje electrónico (e-learning) y de gobierno electrónico (e-bussiness), entre otros muchos. Dichos entornos se caracterizan por contar con la necesidad de disponer de información correctamente organizada, de forma precisa y disponible a tiempo para ser consumida o filtrada por los usuarios. Esta tarea puede suponer el éxito o fracaso de dichos entornos, facilitando la realización de los mismos. En concreto las organizaciones del conocimiento (Knowledge organizations) [7] y sus trabajadores se han visto impulsadas dentro de la incipiente sociedad de la información. Actividades como descripción documental, indexado y clasificado de recursos de información son en la actualidad grandes retos, dado al comentado dominio de datos dinámico y heterogéneo. En esa línea, Hjørland [8] estableció diferentes técnicas como la utilización de un conjunto de vocabularios limitados al dominio, técnicas de recuperación de información, uso de vistas cognitivas enfocadas en los usuarios o de uso de medidas bibliométricas en con el fin de facilitar las tareas de los trabajadores del conocimiento. Como consecuencia, las organizaciones con la posibilidad de gestionar el conocimiento por medio de estas actividades pueden tomar ventajas, descubriendo

nuevas oportunidades de negocio, o posibilidades de trabajo, realizando análisis de sus actividades diarias. En este sentido el campo de visión tecnológica es uno de las actividades principales dentro de este tipos de empresas (por ejemplo la empresa Treelogic S.L., ha centrado una gran porción de su modelo de negocio en este tipo de tareas) en el que los trabajadores añade nuevas noticias, artículos de blogs o revistas científicas, de forma que permita tener un corpus de información con el objetivo de descubrir nuevas oportunidades de negocio, permitiendo a las compañías a estar actualizadas en cuanto a las diferentes tendencias. En este tipo de ambiente, clasificar y aprovechar estos recursos de información conlleva a que las organizaciones que hagan uso de este tipo de medidas, tomar ventaja de este la información subyacente y por tanto, disponer del "know-how" para mejorar su propio rendimiento.

Existen diferentes técnicas, populares y extendidas, con las que se puede tratar de abordar los requisitos que disponen las organizaciones del conocimiento. Estos requisitos se centran en la necesidad de modelar, estructurar, organizar y aprovechar tanto la información como su contenido dentro de dichas organizaciones. Entre estas técnicas se encuentran los mapas conceptuales, las taxonomías o los vocabularios controlados. Sin embargo, de forma reciente, ha habido un gran crecimiento en el uso de ontologías y folksonomías como métodos para gestionar estos grandes volúmenes de información de forma eficiente. Una ontología es habitualmente descrita como una especificación de una conceptualización compartida [9], es una descripción formal de conceptos y sus relaciones envueltas en un dominio de conocimiento para asistir a organizar y construir sistemas basados en el conocimiento. Por otro lado una folksonomía [10] es el resultado de la libre categorización, sin estructura y ni formalidad. Habitualmente están creado por un grupo de usuarios, aplicando "inteligencia colectiva"[11]; formando por usuarios de diferentes niveles de conocimiento interesados en un tema en concreto. Haciendo uso de estos dos métodos en alza, se ha realizado un gran número de desarrollos exitosos dentro de campos tan dispares como la salud electrónica (e-Health), el gobierno electrónico (e-Government) o la tramitación electrónica de contratos públicos (e-Procurement) (ontologías [12][13][14]) así como sitios web 2.0 (folksonomías [15][16][17][18][19]) como Delicious[1], Fickr[2], Youtube[3], Likedin[4], Twitter[5], Facebook[6].

Generalmente, las ontologías son más adecuadas para clasificar información con un dominio específico, donde el tipo de información está restringida por un contexto, en el que expertos se encargan de realizar dicha categorización con un consenso previo entre ellos. En este contexto, los datos, la información y el conocimiento son estáticos y estables. Este tipo de propuesta sufre, de forma evidente, ciertos inconvenientes a la hora de gestionar y organizar información en ambientes dinámicos y cambiantes. Sin embargo ciertas ontologías como DOLCE[7], SUMO[8] y PROTON[9] ha sido desarrolladas con el objetivo de formalizar entidades y sus relaciones, con un alto nivel de abstracción. Por otra parte, como resalto Shirky [20], una folksonomía se adapta mejor a entornos donde el conjunto de datos es grande y dinámico, las entidades son inestables y participan usuarios con diferentes niveles de conocimiento. Estas

[1] https://delicious.com/
[2] http://www.flickr.com/
[3] http://www.youtube.com/
[4] https://www.linkedin.com/
[5] https://twitter.com/
[6] https://www.facebook.com/
[7] http://www.loa.istc.cnr.it/DOLCE.html
[8] http://www.ontologyportal.org/
[9] http://proton.semanticweb.org/

características encajan con los datos provenientes de internet, por tanto el uso de folksonomías se adaptan mejor proveyendo servicios de inteligencia corporativa, de sugerencia publicitaria, o de descubrimiento del conocimiento. Más allá, otro aspecto a tener en cuenta en la gestión de información recae en el comportamiento de los usuarios, trabajos previos [21] han demostrado que una vez una etiqueta ha sido incluida dentro del sistema, esta tiene una probabilidad mayor de volver a ser utilizada de nuevo, obedeciendo al principio de Pareto o regla del 80-20 [22].

1.2 FINALIDAD DEL PROYECTO

De acuerdo con los puntos previamente comentados, el crecimiento de datos e información en formatos digitales, multilingües y dinámicos, hace que las organizaciones del conocimiento hagan uso de ontologías y folksonomías como métodos de gestión. Por este motivo, en esta investigación, presentamos una metodología multilingüe e hibrida que permita mediante técnicas de etiquetado automático, y recomendación de etiquetas basadas en el comportamiento previo de los usuarios con el sistema, construir una folksonomía haciendo uso de técnicas como procesamiento de lenguaje natural, consultas a ontologías y folksonomías externas o técnicas de aprendizaje automático con el objetivo de proporcionar una metodología multilingüe que no sufra de arranque en frío [23]. Sobre esta, se ha desarrollado una librería llamada ACOTA (Automatic Collaborative Tagging) como implementación de referencia de la misma, que provee un motor de etiquetado automático con funcionalidades colaborativas permitiendo hacer recomendaciones basadas en el comportamiento previo de los usuarios.

1.3 PROYECTO ORGANIZACIONES INTELIGENTES GLOBALES INNOVADORAS (ORIGIN)

Figura 1.1 Logo del proyecto ORIGIN y ACOTA

El proyecto, se desarrolla en el período 2010-2013, perteneciente al Fondo Tecnológico, en las regiones de Castilla la Mancha y Asturias. El Fondo Tecnológico es una partida especial de fondos FEDER[10] de la Unión Europea dedicada a la promoción de la I+D+i empresarial en España y está gestionado por el CDTI[11] (Centro para al Desarrollo Tecnológico Industrial) dentro del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008-2011.

[10] http://europa.eu/legislation_summaries/agriculture/general_framework/g24234_es.htm
[11] https://www.cdti.es/

El objetivo del proyecto ORIGIN[12] es aumentar la productividad de las actividades de desarrollo de software en escenarios globales, mejorando la calidad de los productos desarrollados, incrementando de esta forma el nivel competitivo a nivel internacional de las empresas compuestas por factorías de software.

Figura 1.2 Organizaciones Gestoras

Para conseguir este objetivo se desarrollará un conjunto de herramientas conceptuales, metodológicas y sistemas que permitan optimizar la fabricación de software en este tipo de escenarios, paliando los problemas de comunicación y gestión de conocimiento y asegurando la calidad del software producido.

Del mismo modo, las tecnologías desarrolladas facilitarán la utilización del conocimiento organizacional en las actividades de desarrollo, facilitando la toma de decisiones en las actividades de la organización, con el objetivo de incrementar su competitividad, centrándose para ello en la innovación y excelencia como factores diferenciadores.

1.3.1ACOTA dentro de ORIGIN

El proyecto ACOTA ha sido desarrollado como parte de un sub-paquete de trabajo (2.7 Sistema para el soporte de la excelencia organizacional) dentro del proyecto ORIGIN (Organizaciones Inteligentes Globales Innovadoras).

Figura 1.3 Empresa Colaboradora

Este proyecto ha sido desarrollado en colaboración entre WESO RG [13] (Grupo de investigación Web Semantics Oviedo Research Group perteneciente a la Universidad de Oviedo) y Treelogic S.L. [14] con el cometido de ser integrado dentro de la herramienta de alerta dentro su

[12] http://innovation-labs.com/origin/
[13] http://www.weso.es/
[14] http://www.treelogic.com/

herramienta de vigilancia estratégica en una empresa, un centro investigador o un departamento de I+D+i.

Capítulo 2. FIJACIÓN DE OBJETIVOS

El trabajo de investigación *ACOTA: TECNOLOGÍAS DE ETIQUETADO SEMIAUTOMÁTICO Y COLABORATIVO* tiene como objetivo el desarrollo de una metodología multilingüe semiautomática y colaborativa para la etiquetación de documentos. Dicha metodología trata de reducir el esfuerzo del acto de etiquetación, por parte de los usuarios. Resumiendo, el principal objetivo de la presente investigación es:

> Desarrollo de una metodología multilingüe semiautomática y colaborativa para la etiquetación de documentos.

Dentro de la investigación, también se deben alcanzar otros objetivos secundarios como:

1. Estudiar el estado del arte en sistemas de recomendación, de etiquetación, de recomendación, así como sistemas afines que puedan ser aprovechados dentro del marco de esta investigación

2. Identificar el conjunto de tecnologías y técnicas que puedan ser aprovechadas para el desarrollo de tanto la metodología como su implementación de referencia.

3. Diseñar una metodología multilingüe semiautomática y colaborativa para la etiquetación de documentos

4. Diseñar y construir un prototipo que implemente la metodología propuesta.

5. Evaluar la validez de la propuesta, así como divulgar científicamente los resultados del presente trabajo mediante publicaciones a revistar internacionales y/o congresos.

2.1 POSIBLES ÁMBITOS DE APLICACIÓN

El presente proyecto se centra en el desarrollo de una metodología multilingüe semiautomática y colaborativa para la etiquetación de documentos (Sitios Web, Artículos, Libros, etcétera). Uno de los objetivos principales de este proyecto es reducir el esfuerzo asociado a la etiquetación de documentos. El producto obtenido mediante la presente investigación tendría como posibles ámbitos de aplicación:

2.1.1Gestión del Conocimiento empresarial

El concepto de gestión del conocimiento está íntimamente ligado a la gestión de las condiciones, del entorno, y todo lo que hace posible que el conocimiento se cree, almacene y transmita dentro de una organización [24].

En la actualidad el manejo de la información y del conocimiento lo es todo. En el ámbito empresarial esta realidad es mucho más latente: la correcta administración de la información y de los conocimientos del mercado es de vital importancia y tiene gran repercusión económica.

Todos los elementos que ayuden a la gestión de la información y el conocimiento empresarial revierten en una mejora de la eficacia de los métodos utilizados, lo que a su vez trae consigo una mejora de las ganancias económicas. La aplicación de componentes que implementen la metodología propuesta en este proyecto de investigación podría mejorar la capacidad de las empresas para la gestión del conocimiento de sus procedimientos y de su entorno, competencia y actualizaciones tecnológicas en general.

2.1.2Extracción y Recuperación de Información

El número y tamaño de las fuentes de información están en constante aumento, los sistemas e extracción y recuperación de información se centran en identificar y extraer información de las mencionadas fuentes. La aplicación de la presente metodología permite categorizar y etiquetar la información según se va extrayendo, facilitando la creación de metadatos para el consumo por parte de diferentes servicios informáticos.

Entre las ventajas ofrecidas se pueden nombrar: la especialización de los resultados de búsqueda, capacidad de sugerencia de información relacionada, adaptación al contexto de la búsqueda y clasificación de los resultados según relevancia o tipo.

2.1.3Sistemas E-Learning

Los sistemas de aprendizaje electrónico son utilizados tanto para consumir como para producir información. Disponer de estos contenidos organizados es un aspecto crucial para el desarrollo éxitos de los mismos. En los últimos años los centros educativos y más recientemente los *Massive Online Open Course (MOOCS)* están centrándose en soportar educación multilingüe. Por estos motivos, los sistemas de aprendizaje electrónico son un ámbito excelente para la implantación de un sistema multilingüe semiautomático y colaborativo.

La implantación de dicho sistema permitiría disponer de toda la información etiquetada de forma que facilite la búsqueda de la misma, así como permitir filtros para acceder a la misma.

2.1.4Sistemas E- Government

Los sistemas de gobierno electrónico se centran en la transparencia y la interactividad [25]. Este tipo de sistemas producen una gran cantidad de información, dicha información puede estar en diversos idiomas, teniendo en cuenta que existen un gran número de países que cuentan más de una lengua oficial, por lo tanto permitir la etiquetación de contenidos multilingües es un requisito esencial para el correcto desempeño de este tipo de sistemas.

La implantación de la metodología propuesta en el ámbito de sistemas E-Gobernment permitiría favorecer la interactividad, al disponer de toda la información etiquetada y accesible para los usuarios de dicho sistema.

2.1.5Ambientes B2B & B2C

La globalización de la información ha permitido relaciones entre personas y/o empresas que en otros tiempos no habrían sido posibles. En este punto, la interacción entre empresas ha

traído como consecuencia la necesidad de la creación de sistemas de intercambio de información. Disponer de un sistema que "comprenda" los mismos términos, facilita en gran medida dicho intercambio.

Por consiguiente, la implementación de una folksonomía creada a través de tecnologías colaborativas, podría repercutir en la creación de fuentes fiables de términos y procesos, permitiendo un mejor entendimiento entre empresas y consumidores. Además, estas analogías podrían facilitar la interactividad entre diferentes sistemas entre dos o más actores como pueden ser empresas, consumidores y organismos públicos.

2.1.6CMSs

Los sistemas de gestión de contenidos son utilizados por organizaciones que requieren publicar grandes cantidades de contenido a Internet. Dicho contenido en función del contexto, puede estar una variedad de idiomas. Por tanto gracias a la aplicación de nuestra metodología se puede etiquetar dicho contenido, facilitando tanto el consumo como la producción de contenidos.

Por lo tanto podría ser aplicado en diferentes contextos, como puede ser los blogs, foros, wikis, sistemas de comercio electrónico, sistemas de publicaciones digitales entre un largo etcétera.

Capítulo 3. ESTADO ACTUAL DE LOS CONOCIMIENTOS CIENTÍFICO-TÉCNICOS

3.1 ESTADO DEL ARTE

3.1.1Medios para Categorizar Información

El hilo conductor de esta investigación está basado en la necesidad de organizar la información para su posterior explotación en diversas áreas del interés humano, como lo son la educación, las ciencias, la economía y el entretenimiento. Ésta necesidad de organización y clasificación de recursos e información no es una nueva; como se plantea en [26], se han ido construyendo distintas herramientas, basadas en mayor o menor medida en representaciones conceptuales reducidas directamente de la representación del conocimiento plasmada en los documentos.

Como ejemplos de estas representaciones se presentan conceptos como:

•Listas: relaciones de equivalencia o enumeración [27].

•Tesaurios: estructuras que describen relaciones de equivalencia, jerarquía y asociativas entre términos o conceptos [27].Según [26] son estructuras usadas desde la década de los 50's.

•Taxonomías: El término procede del griego taxis, ordenación y nomos, norma. El primero en utilizar el término fue Aristóteles, al categorizar objetos científicos [26]. Una taxonomía es el resultado de clasificar de acuerdo con unas características comunes. Las primeras taxonomías sirvieron para clasificar seres vivos desde la botánica o la biología, principalmente a partir del XVIII. En los últimos años, las taxonomías se han venido usando en sitios Web y en entornos corporativos para la organización de su información [26]. En resumen, son entonces estructuras que definen relaciones de jerarquía entre términos u conceptos [27].

•Folksonomía: estructuras de clasificación de recursos e información logradas a través del etiquetado de información a través de palabras en lenguaje natural en entornos sociales digitales [26].

•Ontología: estructura de representación del conocimiento que permite tanto la organización de conocimiento como su reutilización y la inferencia de otro nuevo [26].

Tomando como vista general de ámbito de la investigación las herramientas propias del ambiente digital y en particular las asociadas con los entornos web, se pasa a la descripción más detallada de ontologías, folksonomías y herramientas para el logro de diversas estructuras taxonómicas como lo son las herramientas de etiquetado de recursos e información.

3.1.2Ontología

El término ontología proviene de un término filosófico, dónde la ontología es la disciplina que indaga en los problemas del ser y de la realidad [28], la discusión sobre la existencia, la representación de lo real y todos sus problemas derivados son abordados en esta disciplina. Tal y como es planteado en [29], en un sentido filosófico, una ontologías es un sistema particular de las categorías que representan una cierta visión del mundo.

Dentro del ámbito de la ingeniería, y en concreto dentro de la Inteligencia Artificial (AI), una ontología es un artefacto de ingeniería, construido con un vocabulario específico utilizado para describir una determinada realidad, formado por entidades y sus relaciones, además de un conjunto de supuestos explícitos sobre el significado de las palabras de un vocabulario. La definición más aceptada en el ámbito de la ingeniería es la de [9] que describe una ontología como una especificación de una conceptualización compartida, es decir, una descripción formal de los conceptos y relaciones que intervienen en un dominio. Una definición algo intangible pero lo suficientemente general para englobar todos los aspectos que se quieren manejar dentro del concepto.

3.1.3Folksonomía

La gran cantidad de información y recursos ofrecida a través de internet en las últimas décadas, así como el advenimiento de la "Web 2.0"[30], propició el nacimiento de plataformas online que permiten gestionar información sobre diferentes campos basadas en herramientas de etiquetado tal y como se comentó en la sección anterior (IMDB[15] para cine, LastFM[16] para música o Flicker[17] para imágenes y Delicious[18] para marcadores, son ejemplos de este tipos de sitios). Este tipos de sistemas se caracteriza por hacer uso de clasificación mediante técnicas de etiquetado, estas son realizadas por cualquier usuario independientemente de su preparación, o conocimiento dentro de la materia. Thomas Vander Wal realizó diversos estudios sobre este nuevo fenómeno, acuñando un nuevo término para definir este tipos de ténica de categorización, la Folksonomía, que proviene de la unión de las palabras "folk" que significa popular o folklórico y "taxonomía" que significa clasificación [10].

Tal y como afirma Vander Wal [10], una folksonomía es el resultado de etiquetado libre de la información. Barrueta [31] en cambio define una folksonomía como una taxonomía abierta, generada y ampliada por sus usuarios, y cuyos elementos se utilizan para etiquetar recursos.

3.1.4Ontología Vs Folksonomía

Los ámbitos de aplicación de las folksonomías y las ontologías ha sido descrito por Shikly [20], ver Tabla 3.1, como puede observarse las ontologías se centran en ambientes formales y estables, en los que las posibilidades de cambio sean mínimas y dicha categorización sea

[15] http://www.imdb.com/
[16] http://www.lastfm.es/
[17] http://www.flickr.com/
[18] https://delicious.com/

realizada por un grupo de usuarios expertos. *Caso de Uso:* Poniendo un ejemplo de ontologías en el mundo real, el proyecto europeo Coglaboration[19], actualmente en desarrollo, tiene entre sus tareas el desarrollo de una ontología de los diferentes tipos de agarres que pueden ser realizados por una mano humana (ámbito estable, sin cambio, limitado, formalizable), utilizando como catalogadores a expertos en Psicología y medicina (Expertos, Con consenso, autoritarios) de la universidad de la Universidad de Birmingham[20].

Por el contrario, las folksonomías suelen ser utilizadas cuando el ámbito es más extenso e inestable, soliendo utilizando como catalogadores a usuarios con un conocimiento dispar sobre la materia. *Caso de Uso:* Poniendo como ejemplo Delicious, un sistema donde ser permite compartir marcadores provenientes de Internet (Corpus Inestable, No Formalizable, Variable…), es categorizado por usuarios con el único requisito de crearse una cuenta en el servicio (Amateur, Nóveles, Sin Organización, Falta de Consenso) para comenzar a categorizar dicha información.

Tabla 3.1Tabla Comparativa entre Ontologías y Folksonomías

Características	**Ontologías**	**Folksonomías**
Corpus	Pequeño	Grande
Categorías	Formales	No Formales
Entidades	“Estables”	“Inestables”
Entidades	“Limitadas”	“Ilimitadas”
Clear Edges	Clear Edges	No Clear Edges
“Catalogadores” expertos	“Catalogadores” expertos	Usuarios desorganizados
Tipo de Usuarios	Autoritarios	Amateurs
Usuarios	Coordinados	Nóveles
Usuarios	Expertos	Sin Autoridad

3.1.5Procesamiento de Lenguaje Natural

Tal y como ha sido descrito por *Vallez et al.* [32]El "Procesamiento del Lenguage Natural" (NLP) es una disciplina con una larga trayectoria con sus orígenes en la década de 1960, como un subárea a medio camino entre la Inteligencia Artificial y la Lingüística, con el objeto de estudiar los problemas derivados de la generación y comprensión automática del lenguaje natural. Dentro de este tipos de técnicas existen técnicas enfocadas la detección lingüística, detección morfológica, o procesamiento estadístico del lenguaje natural.

3.1.5.1 LanguageDetector

El presente trabajo tiene entre sus objetivos en realizar una metodología multilingüe, por lo tanto existe la necesidad de disponer algún tipo de técnica que permita determinar el idioma del texto que se está etiquetando. Pese a que existen diferentes técnicas y herramientas que permitan realizar esta tipo de tarea, hemos utilizado la técnica de categorización basada en

[19] http://www.coglaboration.eu/
[20] http://www.birmingham.ac.uk/index.aspx

detección de n-gramas [33] por medio de técnicas de análisis bayesiano debido a su simpleza y precisión.

Figura 3.1 Diversidad Multilingüe

Hemos utilizado la librería language-detection[21], desarrollado por Shuyo Nakatani y licenciada bajo licencia Apache 2.0, al ser una implementación de referencia de dicha investigación, así como permitir la detección de una gran cantidad de idiomas (53 idiomas en la última versión), así como ser fácilmente extensible por medio de plugins.

3.1.5.2 Apache Lucene

Figura 3.2 Logo de Apache Lucene

Apache Lucene es una librería de alto rendimiento dedicada a tareas de búsqueda de texto, ha sido desarrollada por la fundación Apache y está licenciada bajo licencia Apache 2.0, resultando una librería escalable generalmente utilizada para realizar funciones minería de texto, procesamiento de texto y recuperación de Información (IR) Recuperación de la información es definido por *Baeza Yates et al.* como [34] el al proceso de búsqueda de documentos, búsqueda de información a partir de los metadatos. Gracias a la implementación de esta librería, es posible añadir capacidades de búsqueda a diversas aplicaciones

3.1.5.3 Apache OpenNLP

Figura 3.3 OpenNLP Logo

Apache OpenNLP[22] es una máquina de aprendizaje basada en un kit de herramientas para el procesamiento de lenguaje natural. Ha sido desarrollado por la fundación Apache y está

[21] https://code.google.com/p/language-detection/
[22] http://opennlp.apache.org/

licenciada bajo licencia Apache 2.0, surgiendo como parte de los proyectos "incubator"[23]. Soporta la gran mayoría de las tareas más sencillas, como puede ser tokenización, detección de oraciones, etiquetado morfosintáctico, extracción de entidades nombradas, chunking, parsing, y resolución de correferencias. Estas tareas son usualmente requeridas para crear servicios de procesamiento de texto más avanzados.

El objetivo del proyecto OpenNLP es el de la creación de un conjunto de herramientas maduras para las tareas antes mencionadas. Un objetivo adicional es proporcionar modelos pre-construidos para una variedad de idiomas, así como los recursos de etiquetado de texto que esos modelos se derivan [35].

3.1.6Aprendizaje Automático

El Aprendizaje Automático (Machine Learning) tal y como define en libro *Machine Learning* [36] es una rama de la Inteligencia Artificial que tiene como objetivo el desarrollo de técnicas que permitan a las sistemas informáticos la capacidad de aprendizaje. Entiéndase aprendizaje como la capacidad de generar comportamientos a partir de una información, generalmente, no estructurada. Este tipo de información puede ser de tres tipos, en función como haga uso de la misma, de crecimiento, de reestructuración y de ajuste. Este tipo de técnicas son habitualmente utilizadas en sistemas de recomendaciones así como en técnicas de minería de datos.

3.1.6.1 Mahout

Figura 3.4Logo de Apache Mahout

Apache Mahout[24] es un proyecto desarrollado por la fundación de Apache licenciado bajo licencia Apache 2.0. Surge como una evolución de un subproyecto del proyecto Apache Lucene. Mahout (cuyo significado es "Cuidador de elefantes) es un motor de recomendaciones construido sobre construido sobre el framework distribuido (implementación libre del algoritmo [37] de Google) Apache Hadoop (cuyo logo es un elefante), de ahí el juego de palabras.

Las principales funcionalidades de esta librería s pueden enumerar son enumeradas por [38] por *Owner et al.*:

•Desarrollo de aplicaciones inteligentes

[23] http://incubator.apache.org/
[24] http://mahout.apache.org/

•Para el análisis de problemas relacionados con el área de inteligencia artificial y aprendizaje automático

•Manejo grandes volúmenes de datos sin la necesidad de grandes recursos tecnológicos

3.2 ANTECEDENTES

En el ámbito del tratamiento, almacenamiento, extracción y recuperación de información en internet son muchos los estudios realizados. Este proyecto de investigación se enfoca en los que buscan respuestas a éstos problemas aplicando los conceptos antes mencionados, como la utilización de folksonomías, consultas a ontologías y folksonomías, técnicas de etiquetado y de análisis de lenguajes natural y etiquetado automático

3.2.1Improved annotation of the blogopshere via autotagging and hierarchical clustering

Con el objetivo de reducir el esfuerzo de etiquetar artículos de blogs, en *Improved annotation of the blogopshere via autotagging and hierarchical clustering* [39] se describe un sistema que hace uso del top-tres de un raninkg frequency-inverse document frequency (TFIDF[40]) de etiquetas, con el objetivo de agrupar documentos dentro de cústeres en función del número de etiquetas compartidas por los mismos.

Esta técnica agrupa conjuntos de documentos con la hipótesis de que aquellos documentos que están dentro de un mismo clúster tienen más posibilidades de ser similares que aquellos que no están en diferentes clústeres.

Partiendo de un conjunto de datos en inglés extraídos del motor de búsqueda de blogs Technorati[25], se realizó los cálculos pertinentes para determinar la similitud de documentos y generar el ranking TFIDF, obteniendo buenos resultados. *Brooks et al*. dejan abierta la posibilidad como trabajo futuro, de realizar análisis de las frases y de realizar búsquedas de sinónimos con el fin de mejorar los resultados del sistema.

3.2.2AutoTag

Mishne et al. evolucionaron las técnicas presentadas por *Brooks et al.* centrándose en el componente social y/o colaborativo. Este trabajo descrito en *AutoTag: A Collaborative Approach to Automated Tag Assignment forWeblog Posts* [41] presenta AutoTag, un sistema de sugerencia de etiquetas para artículos de blogs (blog-posts) que hace uso de métodos de filtrado colaborativo. Esta herramienta ofrece recomendación de etiquetas basadas, basadas en el conjutno de etiquetas asociadas a documentos similares, relegando al autor del artículo la tarea de escoger de entre las etiquetas propuestas por el sistema aquellas apropiadas y representativas.

[25] http://technorati.com/

AutoTag hace uso de técnicas de técnicas de recuperación de información para calcular la similitud entre los documentos, y a partir de la misma poder realizar las diferentes recomendaciones de etiquetas. La herramienta propuesta por *Mishne et al.* obtienen buenos resultados, permitiendo reducir el esfuerzo de la tarea de etiquetación y por tanto ayudando a bloggers, e indirectamente a los usuarios que podrán hacer uso de las susodichas etiquetas.

3.2.3TagAssist: Automatic Tag Suggestion for Blog Posts

TagAssist es otro sistema de recomendación de etiquetas para artículos de blogs, descrito en *TagAssist: Automatic Tag Suggestion for Blog Posts* [42], resalta sobre los trabajos anteriores en añadir soporte para compresión de etiquetas (para el idioma inglés), una técnica que consigue reducir la ambigüedad de términos dentro del sistema, permitiendo incrementar la precisión del mismo de forma notoria. Dicho sistema está formado por dos pasos principales, en primer lugar se realiza extracción de información, formándose un ranking de etiquetas y a continuación se procede a realizar la compresión de las diferentes etiquetas.

En esta investigación llegan a la conclusión de que según se va aumenten el tamaño del corpus, y por tanto de la folksonomía, los resultados de este sistema se ven incrementados, gracias a disponer de una mayor cantidad de información, sobre los que realizar estas técnicas.

3.2.4A language model approach for tag recommendation

Sun et al. presenta en *A language model approach for tag recommendation* [43], una evolución del trabajo desarrollado por *Brooks et al.* y *Song et al* [44]. Esa investigación evoluciona dichas técnicas añadiendo soporte para recomendación de etiquetas basadas en modelos de lenguaje (LMTR), esta aproximación consiste en utilizar como medida de similitud de documentos el conjunto de términos que tienen en común dichos documentos, aplicando una serie de técnicas estadísticas con ese fin.

Este trabajo es evaluado contra un conjunto de documentos provenientes de la red social de referencias científicas CityULike, así como la red social de preguntas y respuestas china (Q&A) Zhishi[26], obteniendo un gran rendimiento dado al bajo cantidad de tiempo invertido por cada documento.

3.2.5Automatic keyphrase Extraction extraction based on NLP Automatic Keyphrase based on NLP and statistical methods and Statistical Methods

En el artículo Automatic Keyphrase Extraction based on NLP and Statistical Methods[45] *Dosta et al.* presentan una técnica de recomendación de etiquetas basado en el aplicando procesamiento de lenguaje natural así como el uso de técnicas estadísticas como Term frequency – Inverse document frequence (TF-IDF).

[26] Esta servicio parece haber sido discontinuado por Baidu

En primer lugar esta herramienta se encarga de extraer el texto de los diferentes documentos, a continuación proceden a tokenizar el texto y a partir de los distintos tokens, calculando, se calculan las etiquetas POS (Part of Speech) equivalentes. A partir de dichas etiquetas y teniendo en cuenta la categoría gramatical de cada conjunto de tokens, se procede a extraer tanto unigramas como n-gramas, generándose el ranking TF-IDF a partir del conjunto de etiquetas representativas producidas por el componente automático.

3.2.6Emergent Ontologies by Collaborative Tagging for Knowledge Management

Centrándose en recomendación de etiquetas basado en componentes semánticos, *Jimenez et al.* han desarrollo una metodología para la creación y mantenimiento de ontologías. Dicha metodología es descrita en *Emergent Ontologies by Collaborative Tagging for Knowledge Management* [46]. Este trabajo, hace uso de diversas técnicas para realizar etiquetado de forma automática. De entre las etiquetas propuestas, aquellas elegidas por los usuarios son almacenados en formatos semánticos (RDF[27] y OWL[28],), conformando una ontología. Dicha ontología pueda ser consumida tanto por personas como por maquinas mediante un endpoint SPARQL[29].

(a) Integración de herramientas de etiquetado.

(b) Almacenamiento de etiquetas en lenguajes accesibles por entidades informáticas.

(c) Integración de un sistema de recomendación de etiquetas o términos.

(d) Integración de sistemas de muestra de etiquetas más populares.

(e) Identificación de términos existentes y relacionados.

(f) Recomendación de relaciones entre términos nuevos y términos existentes.

(g) Integración de sistemas de definición de relaciones.

3.2.7Resumen

La gran mayoría de los trabajos previamente mencionados hacen uso prácticamente exclusivo del inglés. En el caso de *Brooks et al.* [39] y de TagAssist [42] eliminan todo el contenido que no está en este idioma (En el caso de este último trabajo, se permiten el uso de una pequeña lista blanca de términos…). Por otro lado, AutoTag [41] mantiene una aproximación menos agresiva manteniendo contenido en idiomas distintos al inglés aunque obtiene bajos resultados para este tipos de contenidos al hacer uso de un steamer pensado para ser utilizado con contenido en inglés. Cabe resaltar que tal y como ha sido mencionado en secciones anteriores,

[27] http://www.citeulike.org/
[28] http://www.w3.org/RDF/
[29] http://www.w3.org/TR/rdf-sparql-query/

debido al creciente uso de información multilingüe, existe una necesidad real de aportar sistemas de recomendado que soporten dicho tipo de contenido.

Otro aspecto a tener en cuenta es el de arranque en frío [23]. Generalmente, los sistemas de recomendación con apenas uso (y por tanto dispone de un corpus pequeño), no son capaces de realizar recomendaciones adecuadas debido a la falta de información mínima para poder realizar dichas recomendaciones. Por tanto una característica esencial que debería de disponer un sistema de recomendación hibrido es la posibilidad de minimizar este tipo de problema mediante el uso de sistema de recomendación automático que funcione aun cuando el sistema esté frio.

A partir de este conjunto de trabajos previos hemos extraído una serie de características que una metodología multilingüe semiautomática y colaborativa para la etiquetación de documentos debería de tener, ver Tabla 3.1.

Tabla 3.2 Características de un Recomendador de Etiquetas Multilingüe semiautomático y colaborativo

Característica
Multilingüe
Colaborativo
Sin Dominio Específico
No sufre de arranque en frío (No Requiere de Entrenamiento)
Técnicas de Procesamiento de Lenguaje Natural
Compresión de Etiquetas
Consultas a Ontologías
Consultas a Folksonomías

Capítulo 4. DESCRIPCIÓN DE LA METODOLOGÍA PROPUESTA

En esta sección mostramos una descripción detallada de la metodología propuesta. Esta metodología, y por ende ACOTA, está dividida en dos etapas principales: 1) el sistema de etiquetado automático se encarga de normalización de texto, extracción de palabras clave y enriquecimiento de etiquetas. 2) La recomendación de etiquetas que utiliza como entrada la salida de la primera etapa recomienda etiquetas basado tanto en el contexto como en el comportamiento previo de los usuarios con el sistema.

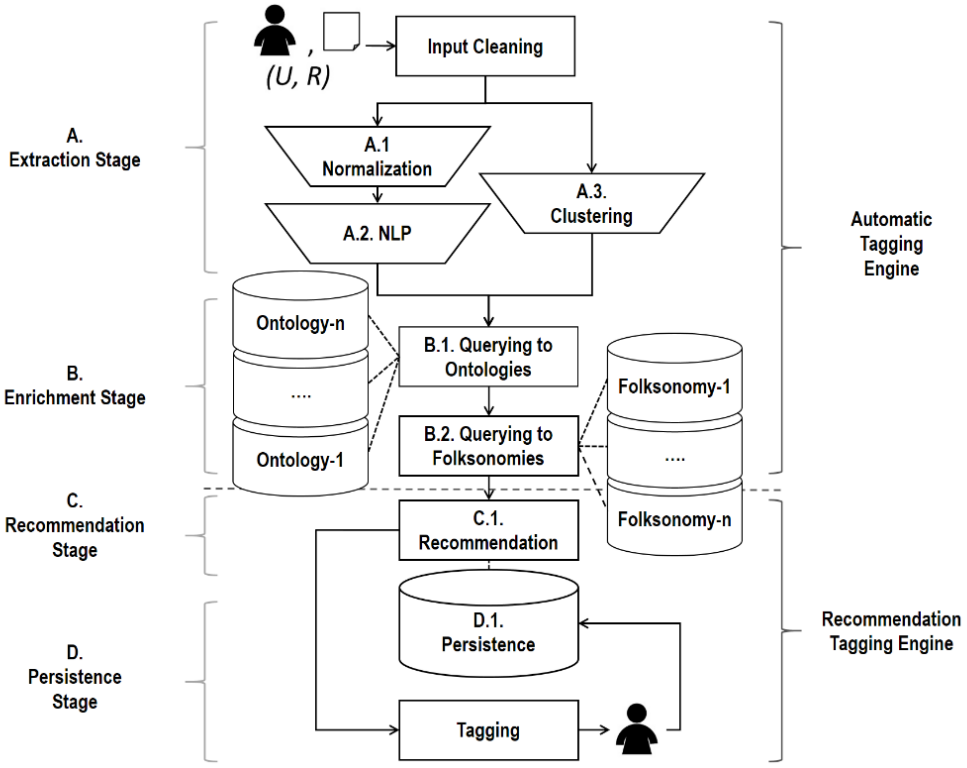


Figura 4.1 Metodología multilingüe semiautomática y colaborativa para la etiquetación de documentos

4.1 MOTOR DE ETIQUETADO AUTOMÁTICO

El motor de etiquetado automático está compuesto por dos etapas. En primer lugar se ejecuta la etapa de extracción, que se encarga de extraer el contenido de los documentos produciendo un conjunto de tokens formados tanto por unigramas como por n-gramas. A continuación se ejecuta la etapa de enriquecimiento, que consiste en realizar consultas tanto a ontologías como a folksonomías con el objetivo de enriquecer los resultados de la etapa anterior. Como resultado se obtiene un conjunto de etiquetas descriptivas para el documento propuesto. Antes de ejecutarse las etapas principales, existe una etapa previa opcional que se

encarga de limpiar la entrada del sistema. Esta etapa previa ayuda a reducir potencial-ruido que puede propagarse de estas primeras etapas, generando etiquetas no descriptivas y por tanto sin utilidad para el sistema.

4.1.1A. Etapa de Extracción

Esta etapa tiene el cometido de extraer etiquetas a partir de un documento. Está compuesto por dos técnicas de extracción diferentes: 1) La extracción de unigramas, se ejecuta mediante la combinación A.1 Normalización y A.2 Procesamiento de Lenguaje Natural, y 2) Extracción de n-gramas, que se ejecutada por una versión refinada de las dos técnicas previamente presentadas.

Figura 4.2 Ejemplo paso a paso de una recomendación

4.1.1.1 A.1 Normalización

Esta fase toma como entrada un recurso y produce un vector de unigramas. El primer paso consiste en partir grupos de palabras por signos de puntuación y a continuación eliminar dichos delimitadores. Cada token se pasa a minúsculas y continuación se eliminan aquellos tokens que tienen menos de tres caracteres por ser habitualmente no relevantes (pronombres, artículos, preposiciones…) para esta fase [45][47]. Para finalizar, se filtran los tokens eliminando

aquellos que están contenido en el grupo de palabras vacías (Este grupo de palabras vacías es dependiente del idioma a procesar, y puede ser configurado), produciendo un conjunto de unigramas.

Por cada unigrama se cuenta el número de ocurrencias y se multiplica por un peso dado. Dependiendo de donde esté ubicado dicho unigrama, el peso es calculado. Dado que las etiquetas que aparecen en el titulo [48] son más relevantes para los usuarios y por tanto disponen de un peso mayor que las que aparecen en el cuerpo.

4.1.1.2 A.2 Procesamiento de lenguaje natural (NLP)

Como *Bohemia et al.* [49] indicó, las etiquetas clave formadas por unigramas suelen estar formado por nombres, adjetivos o participios de verbos (para el inglés). Esta técnica consiste en filtrar la entrada generada por la etapa previa (*A.1*), modificando el peso de la etiqueta basado en el tipo morfosintáctico de la misma.

En primer lugar para cada etiqueta se calcula su categoría gramatical generando una etiqueta POS (Part of Speech) asociada. Adjetivos, nombres y participios son almacenados en función de su categoría gramatical. Las etiquetas que están dentro de la lista negra de tipos gramaticales son eliminadas (Grupos gramaticales que no son relevantes), y finalmente existen un grupo de etiquetas sin procesar. Los pesos de las etiquetas son modificadas en función del vector en el que han sido almacenadas.

Como salida de este vector, se produce subconjunto de las etiquetas recibidas por la fase *A.1 Normalización*.

4.1.1.3 A.3 Agrupamiento (Clustering)

La combinación de las técnicas de *A.1 Normalización* y A.2 Normalización provee un mecanismo rápido y sencillo para generar etiquetas formadas por una única palabra (unigrama). En ciertos casos, los unigramas no pueden aportar suficiente "semántica" dado al poco significado que puede aportar una sola palabra. Por ejemplo si se quisiese etiquetar un documento perteneciente a la Universidad de Oviedo, se podría etiquetar palabras como *"Universidad"* u *"Oviedo"*, que por sí mismas no son tan descriptivas como la etiqueta *"Universidad de Oviedo"*. Por los tanto con el fin de mejorar los resultados de las recomendaciones que pueden ser generadas por el sistema es necesario añadir soporte para etiquetas formadas por bigramas o incluso frases cortas.

El primer paso consiste en fraccionar el conjunto texto de la entada en frases, eliminando signos de puntuación. A continuación se forman grupos de palabras de 1 a *k* tokens. Por cada grupo de palabras se generan sus etiquetas POS en función de las categorías gramaticales de las mismas. Por cada grupo de palabras, se empieza a analizar las etiquetas POS desde los lados del grupo, hasta encontrar una etiqueta POS válida. En el caso de que no se encuentre ninguna etiqueta válida, el grupo de palabras es eliminado. Para finalizar de la misma forma que en *A.1 Normalización* se cuentan las ocurrencias, multiplicando dicha cantidad por un peso, también dependiente de la ubicación del grupo.

4.1.2B. Etapa de Enriquecimiento

La etapa de enriquecimiento recibe un vector de etiquetas, tanto unigramas como n-gramas, como entrada a partir de la salida de la etapa anterior. En esta etapa las etiquetas se enriquecen mediante consultas a ontologías y a folksonomías.

4.1.2.1 B.1 Consultas a Ontologías

Esta técnica consiste en enriquecer etiquetas realizando consultas al diccionario WordNet [50], versión 3.0, para obtener los sinónimos de las etiquetas buscadas.

Cada etiqueta es buscada dentro de WordNet, en el caso de que la etiqueta esté dentro del diccionario, los sinónimos de la misma son extraídos, en el caso de que no esté dentro, esta etapa es finalizada (para la etiqueta buscada). Por cada sinónimo sugerido es buscado dentro del vector de etiquetas, en el caso de que la búsqueda sea exitosa, se incrementa el peso de la misma, en el caso contrario el sinónimo es añadido al vector de etiquetas con un peso por defecto.

4.1.2.2 B.2 Consultas a Folksonomías

Como se ha presentado en secciones anteriores realizar consultas a folksnomías puede ser una espada de doble filo, por un lado incrementa la cantidad de palabras representativos, pero por el otro lado puede incrementar la cantidad de ruido dentro del sistema con bastante facilidad. Teniendo en cuenta todos estos puntos hemos decidido añadir a nuestra metodología una serie de filtros en cadena de forma que solo sean enriquecidas etiquetas sobre un percentil dado (hemos fijado ese percentil a 60, de forma que el 50% de las etiquetas menos representativas no sean enriquecidas por la presente técnica). Por tanto, solo son enriquecidas palabras representativas, saltándose aquellas que no aportan un significado extra. Este proceso de filtrado de forma indirecta ayuda a reducir el número de consultas a servicios externos, generalmente consultas REST[51].

4.1.2.2.1 Mejorando el Rendimiento: Cacheando las consultas REST

Un problema subyacente de la aproximación comentada en esta sección es el consumo de tiempo cuando se necesita enriquecer una etiqueta mediante consultas a servicios externos. Dado que el enriquecimiento de una etiqueta requiere al menos consumir un servicio REST, cuando el número de etiquetas es suficientemente grande, la ejecución de esta etapa puede requerir de una cantidad de tiempo excesiva, incluso con los filtros añadidos por nuestra propuesta.

Figura 4.3 Consulta REST Cacheada

Con el cometido de prever esta situación, se ha añadido una caché completamente configurable que permite anticipar y almacenar los resultados de las consultas REST, ver Figura 4.3, de forma que el sistema consigue reducir el número de consultas necesarias, reduciendo ancho de banda y tiempo necesario para ejecutar dicha tarea. Si bien esta medida aporta grandes ventajas, también tiene como contrapunto una pequeña penalización al realizar la consulta cuando la consulta REST no está cacheada, esto sucede como necesidad de consultar la disponibilidad de la caché (1,2), realizar la consulta de la misma (3,4) y almacenar los resultados dentro de la caché (5), como se puede ver en la Figura 4.4

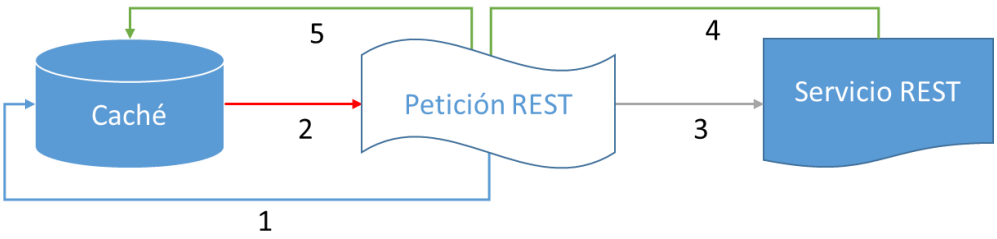


Figura 4.4 Consulta REST sin Cachear

En el caso concreto de sistemas empresariales como pueden ser las intranets, donde la cantidad de documentos procesados es grande, el uso de una caché puede aportar un incremento drástico de rendimiento , dado que una gran cantidad de consultas estarán cacheados, reduciendo el tiempo necesario a realizar una consulta REST a cerca de una petición local. Por otro lado, cachear documentos que no han estén incluidos dentro de la misma conlleva una penalización en el tiempo necesario para realizar el enriquecimiento.

4.2 MOTOR DE RECOMENDACIÓN

El principal objetivo del motor de recomendación es sugerir etiquetas basadas en el comportamiento previo de los usuarios con el sistema.

4.2.1C. Fase de Recomendación

En esta etapa se centra en recomendar etiquetas basadas en el feedback previo de los usuarios con el sistema, para ello se utilizó un algoritmo de recomendación basado en recomendaciones elemento a elemento. Esta técnica de recomendación consiste en recomendar etiquetas basado en como de parecidos son los elementos entre ellos [52], empleando como medida de similaridad el Coeficiente de Tanimoto [53][54]

Explicando el mecanismo de forma sencilla, la idea consiste en si una serie de etiquetas, por ejemplo "*Elvis Presley*", "*Music*", y "*Rock & Roll*" han sido etiquetados a dos documentos, ver Figura 4.5. Por un lado la etiqueta *Music* y *Elvis Preseley* están etiquetados al documento www.elvis.com y por otro lado las etiquetas *Music* y *Rock & Roll* ha sido etiquetados al documento moody_blue. Nuestro sistema dado las relaciones establecidos, es capaz de recomendar etiquetas basadas en estas relaciones subyacentes, de forma que es capaz de inferir nuevas relaciones. De este modo, es capaz de devolver como resultado Elvis Presley ante la etiqueta rock & roll y viceversa.

El principal inconveniente de nuestra propuesta, es cuando el sistema está "frio" [23], dado que el sistema de recomendación no es capaz de realizar recomendaciones . En un sistema de recomendación puro, que solo utilizarán esta técnica para recomendar etiquetas, significaría la incapacidad de recomendar etiquetas y por tanto no sería capaz de realizar su tarea. Nuestra metodología es hibrida, es decir, utiliza un sistema de recomendación automático sobre el que se aplica un sistema de recomendación, por tanto, en el caso de que el sistema estuviera frio, por ejemplo sin ningún dato almacenado, sería capaz de recomendar etiquetas apropiadas, dado que el sistema de recomendación automático seguiría funcionando por debajo, produciendo resultados válidos-.

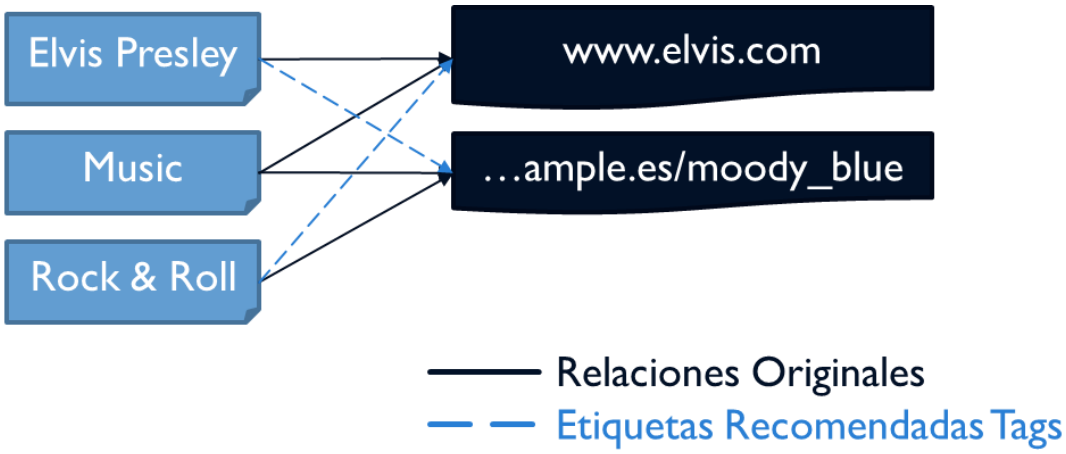


Figura 4.5 Ejemplo Recomendaciones

4.2.1.1 Ejemplo

Por ejemplo la palabra "Java" tiene al menos tres posibles significados, 1) Java el lenguaje de programación de alto nivel orientado a objetos desarrollado originalmente por Sun Mycrosystems 2) Una isla del archipiélago indonesio 3) Tipo de café popular proveniente de la isla de Java con un toque picante y dulce. Estos tres significados difieren de ellos y por tanto en diferentes escenarios el sistema de recomendación debería de comportarse de forma diferente, y adaptado al significado de cada término dentro del contexto del sistema. Por ejemplo un sistema de recomendación de etiquetas dentro de una empresa de desarrollo Software, para el término Java podría recomendar "POO" u "Oracle". En un sistema de recomendación de etiquetas para una gran superficie comercial para el término java podría recomendar etiquetas como "café", "fuerte" o "picante" y por último dentro de un sistema de viajes podría recomendar términos como "isla", "indonesia" o "asia".

4.2.2D. Fase de Persistencia

Una vez las etiquetas generadas por las etapas previas han sido presentadas a los usuarios, estos tienen la opción de seleccionar una de las etiquetas propuestas, o en el caso de que ninguna de las mismas sean de su gusto, puede añadir nuevas etiquetas de su puño y letra. Estas etiquetas, tal y como se comentó en la fase anterior, serán utilizadas para mejorar las recomendaciones realizadas por el sistema.

Capítulo 5. ARQUITECTURA DE ACOTA

Como implementación de la metodología propuesta hemos desarrollado ACOTA (Automatic Collaborative Tagging). ACOTA está compuesto por dos componentes principales, el componente Core, que provee el motor de etiquetado automático y el componente Feedback, que provee las funcionalidades de etiquetado automático y de recomendación. Ambos componentes, tanto Core como Feedback, pueden ser utilizados como proyectos independientes, o ser combinados para tomar ventajas de las funcionalidades que aporta cada componente.

Figura 5.1 Logo de ACOTA

5.1 COMPONENTE CORE

Este componente aporta el motor de etiquetado automático e incluye una serie de Enhancers (Potenciadores) que proveen funcionalidades como procesamiento extracción de palabras clave, de lenguaje natural, o de consultas a ontologías y folksonomías.

Tabla 5.1Relación Entre la metodología propuesta y ACOTA

Fase	Etapa	ID	Componente ACOTA
Motor de Etiquetado Automático	Extracción	A.1	Core
Motor de Etiquetado Automático	Extracción	A.2	Core
Motor de Etiquetado Automático	Extracción	A.3	Core
Motor de Etiquetado Automático	Enriquecimiento	B.1	Core
Motor de Etiquetado Automático	Enriquecimiento	B.2	Core
Motor de Recomendación de Etiquetas	Recomendación	C.1	Feedback
Motor de Recomendación de Etiquetas	Persistencia	D.1	Feedback

5.1.1A. Etapa de Extracción

La etapa de extracción es realizada por la combinación de LuceneEnhancer, OpenNLPEnhancer y OpenNLPEnhancer, implementando *A.1 Normalización*, *A2. Procesamiento de lenguaje na*tural y *A.3 Agrupamiento*, respectivamente.

5.1.1.1 A.1 Normalización

LuceneEnhancer implementa la etapa de normalización, extrayendo unigramas desde un recurso, normalizando los resultados. Este potenciador se implementó utilizando la popular librería para extracción de información *Apache Lucene*[30], utilizándola para procesar la entrada, eliminando palabras vacías, puntuación, etcétera.

5.1.1.2 A.2 Procesamiento de Lenguaje Natural

La etapa de procesamiento de lenguaje natural se realiza por el potenciador *OpenNLPEnhancer*. Este potenciador utiliza la librería *Apache OpenNLP* [31] para realizar la traducción de palabras a etiquetas POS (Part of Speech). Estas etiquetas son utilizadas para realizar las técnicas de filtrado en función del tipo de categoría gramatical de cada etiqueta.

5.1.1.3 A.3 Agrupación

TokenizerEnhancer implementa la fase A.3 Agrupación, emplea una versión refinada de las técnicas anteriormente mentadas, con la finalidad de extraer de n-gramas. TokenizerEnhancer utiliza la librería *Apache OpenNLP* para realizar la traducción de palabras a etiquetas POS.

5.1.2B. Etapa de Enriquecimiento

Esta etapa se centra en realizar consultas a ontologías y a folksonomías, con el cometido de enriquecer el conjunto de etiquetas proveniente de la etapa anterior. Esta etapa está compuesta por dos potenciadores, uno basado en Wordnet (*WordnetEnhancer*) y otro basado en Google Complete API (*GoogleEnhancer*).

5.1.2.1 B.1 Consultas a Ontologías

Esta etapa se realizada mediante el potenciador *WordnetEnhancer*, requiere del diccionario WordNet con el fin de proveer de sinónimos para etiquetas en inglés. Emplea las funcionalidades de la librería Java Wordnet Enhancer [32] (JWI) para cargar el diccionario directamente desde los ficheros de WordNet a memoria, de forma que las consultas se realizan de forma aislada, sin requerir tener WordNet propiamente instalado. De la misma forma, nuevos

[30] http://lucene.apache.org/core/
[31] http://opennlp.apache.org/
[32] http://projects.csail.mit.edu/jwi/

idiomas pueden ser añadidos simplemente disponiendo de los ficheros con los idiomas a utilizar siempre y cuando utilicen un formato compatible que pueda ser procesado por JWI.

Figura 5.2 Logo de WordNet

5.1.2.2 B.2 Consultas a Folksonomías

Utiliza el servicio Google Complete API[33] como folksonomía, enriqueciendo los resultados con nuevas etiquetas. Esta folksonomía, sugiere el top-10 de búsquedas de Google, para para la palabra buscada (ver Figura 5.3), en nuestro caso la etiqueta a enriquecer.

Figura 5.3 Ejemplo de Google Complete API

Como ha sido explicado con anterioridad en la metodología, el prototipo permite emplear un sistema de caché con el fin de optimizar el proceso reduciendo el número de consultas y por ende, reduciendo el ancho de banda y tiempo en ejecutarse el proceso. Como sistema de caché se ha utilizado Memcached[34], un sistema de caché distribuido de alto rendimiento.

5.2 COMPONENTE FEEDBACK

Este componente se corresponde con el Motor de Recomendación de Etiquetas y está divididos en dos subcomponentes, un motor de recomendación que permite al sistema recomendar etiquetas basados en el comportamiento previo de los usuarios con el sistema, así como de un sistema de persistencia que se encarga de almacenar el feedback de los usuarios, permitiendo mejorar la precisión del sistema de recomendación.

5.2.1C.1 Recomendación de Etiquetas

El potenciador LabelRecommenderEnhancer implementa la etapa C1 Recomendación de Etiquetas. Este potenciador sugiere etiquetas basado en el comportamiento previo de los usuarios con el sistema, utiliza la librería Apache Mahout [35], una librería de aprendizaje automático (Machine Learning) que incluye una serie de algoritmos predefenidos. El prototipo

[33] https://support.google.com/websearch/answer/106230?hl=en
[34] http://memcached.org/
[35] http://mahout.apache.org/

actual utiliza una versión modificada del recomendador elemento a elemento de Mahout, con el objetivo de mejora las recomendaciones presentadas por ACOTA.

5.2.2D.1 Persistencia

La librería de aprendizaje automático (machine learning) requiere de un sistema de persistencia donde consultar el feedback previo de los usuarios con el sistema. En el caso de ACOTA se ha decidido conectar Mahout a un sistema de gestión de bases de datos, conteniendo la tupla (documento, ítem, peso). El acceso al sistema de persistencia ha sido diseñado para ser lo más flexible posible, por ello diferentes DBMS tanto relacionales como NOSQL pueden ser conectados a ACOTA. ACOTA de forma nativa incluye soporte para los DBMS relacionales MySQL[36], MariaDB[37] y PostgreSQL[38], así como para el DBMS NSQL MongoDB[39].

Además se deja abierta la posibilidad de que los usuarios puedan extender ACOTA para soportar otros DMBS o inclusos diferentes sistemas de persistencia (Basados en Memoria, Grafos, etcétera).

Figura 5.4. DBMSs Soportados

5.3 ACOTA

Existe una demo permanente de ACOTA disponible en la plataforma PaaS (Plaforma como Servicio) Heroku[40], ver Figura 5.5, en http://acota.herokuapp.com/ .Esta demo permite generar etiquetas a partir de un título, descripción y URI. Además, esta página web incluye información acerca de la librería java y los servicios REST que conforman ACOTA.

[36] http://www.mysql.com/
[37] https://mariadb.org/
[38] http://www.postgresql.org/
[39] http://www.mongodb.org/
[40] https://www.heroku.com/

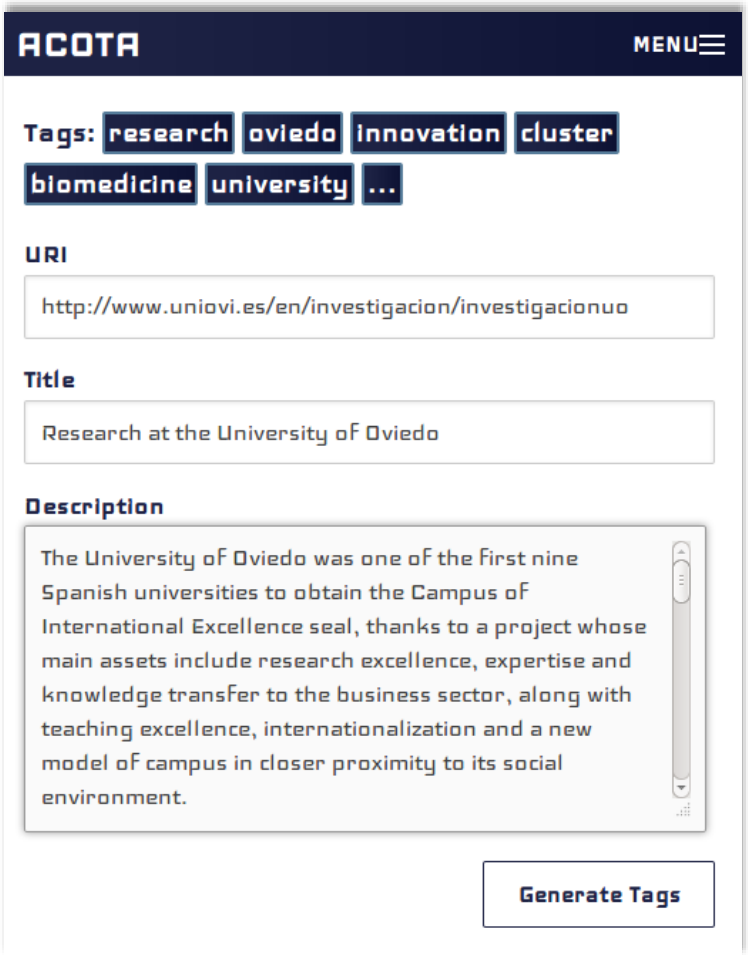


Figura 5.5 Captura de Pantalla de la demo de ACOTA

5.3.1Maven

Uno de los principios que tratamos de seguir con el desarrollo de ACOTA, fue el de simplificar la implantación y uso de la librería. Para ello, todos los componentes de ACOTA utiliza Apache Maven[41], además todos los arquetipos están desplegados en el repositorio o central de Maven[42], de forma que los usuarios que quieran hacer uso de ella solo requerirán de añadir las siguientes dependencias en sus proyectos.

5.3.1.1 Acota-Seed

Repositorio: https://github.com/weso/acota-seed

```
<dependency>
      <groupId>es.weso</groupId>
      <artifactId>acota-seed</artifactId>
      <version>0.3.7</version>
</dependency>
```

[41] http://maven.apache.org/
[42] http://mvnrepository.com/search.html?query=es.weso

5.3.1.2 Acota-Core

Repositorio: https://github.com/weso/acota-core

```
<dependency>
     <groupId>es.weso</groupId>
     <artifactId>acota-core</artifactId>
     <version>0.3.7</version>
</dependency>
```

5.3.1.3 Acota-Feedback

Repositorio: https://github.com/weso/acota-feedback

```
<dependency>
     <groupId>es.weso</groupId>
     <artifactId>acota-feedback</artifactId>
     <version>0.3.7</version>
</dependency>
```

5.3.1.4 Acota-Utils

Repositorio: https://github.com/weso/acota-utils

```
<dependency>
     <groupId>es.weso</groupId>
     <artifactId>acota-utils</artifactId>
     <version>0.3.7</version>
</dependency>
```

Capítulo 6. CASO DE ESTUDIO: GESTIÓN DEL CONOCIMIENTO

Treelogic S.L. es una empresa que proporciona a sus clientes soluciones basadas en las tecnologías de la información y la comunicación. Treelogic S.L. dispone de una solución llamada Imaginn Watching, una herramienta de vigilancia estratégica (alerta tecnológica) para la empresa privada, centro investigador o departamento de I+D+i. El objetivo de este tipo de herramienta consiste en estar actualizado de las últimas tendencias en tecnología y/u oportunidades de negocio dentro del sector donde se desee utilizar.

Figura 6.1Logo Imaginn Watching

El término visión tecnológica consiste según *Davidson et al.* [55] en observar de forma regular diferentes áreas como puede ser la legal, social, tecnológica o medioambiental, con el fin de disponer de información actualizada y lista para ser consultada por los miembros responsables de las diferentes áreas, a la hora de tomar decisiones.

Este tipo de herramienta requiere gestionar y organizar grandes cantidades de información. Esta información suele proveer de Internet, y como fue mentando en secciones anteriores, suele estar formada por una gran variedad de idiomas El dominio de la información cambia constantemente, según nuevos términos, tecnologías y oportunidades de negocio son establecidas [56]. Además, toda esta información debe de poder ser creada y consumida por los trabajadores de la información (Knowledge Workers) de forma colaborativa.

Los requerimientos de una aplicación de alerta tecnología encajan con las características de la metodología multilingüe semiautomática y colaborativa para asistir el conocimiento organizacional definidas en la Tabla 3.2. El soporte multilingüe, la posibilidad de enriquecer los datos con fuentes externas de información, como pueden ser folksonomías u ontologías, la ausencia de dominio específico así como no sufrir arranque en frío y por tanto la falta de necesidad de entrenar el motor de recomendación, están presente dentro de los requisitos que Treelogic S.L. buscaba paliar con el uso de la metodología propuesta de su implementación de referencia, ACOTA.

Como puede ser observado en la Tabla 6.1, ACOTA es comparada con diferentes propuestas (*Brooks et al*, AutoTag, TagAssist y *Bohemia et al.*). No hemos querido evaluar los detalles específicos de cada implementación, y nos hemos enfocado en algunas propiedades generales. Nuestra aproximación encaja con todas las características a excepción de su principal

carencia, la falta de compresión de etiquetas (que ayudaría a mejorar los resultados de nuestra propuesta). Dado que la técnica de compresión de etiquetas está altamente ligado a idiomas en concreto, se decidió desestimar esta opción, pese a que su inclusión mejoraría los resultados de nuestra propuesta. El motivo de esta elección, es el gran esfuerzo que habría supuesto la mejora alcanzable.

Tabla 6.1 Características de un recomendador que soporte técnica multilingüe semiautomática y colaborativas para asistir el conocimiento organizacional

Característica	ACOTA	Brooks	AutoTag	TagAssist	Bohemia
Multilingüe	+	-	-, parcial	-	-
Colaborativo	+	+	+	+	+
Sin Dominio Específico	+	-,Requiere Entrenamiento	-,Requiere Entrenamiento	-,Requiere Rentrenamiento	+
No sufre de arranque en frío (No Requiere de Entrenamiento)	+	-	-	-	-
Técnicas de Procesamiento de Lenguaje Natural	+	-	-	+	+
Compresión de Etiquetas	-	-	-	+	-
Consultas a Ontologías	+	-	-	-	-
Consultas a Folksonomías	+	-	-	-	-

Capítulo 7. METODOLOGÍA DE TRABAJO

A lo largo de este documento se ha presentado un proyecto en el que propone una metodología multilingüe semiautomática y colaborativa para la etiquetación de documentos. En esta sección se presenta la metodología experimental con la que se va a realizar los experimentos para determinar la validez de la metodología propuesta.

7.1 DISEÑO DEL EXPERIMENTO

Para evaluar los resultados se ha decidido utilizar dos métricas diferentes, precisión y recall, ambos propuestos por *Cleverdon et al.* [57].

Precisión (1) es la fracción de etiquetas devueltas por el sistema que son relevantes. En el caso de este experimento, la fracción de etiquetas propuestas que encajan con las seleccionadas por los usuarios reales.

$$Precision = \frac{Match}{|\{Recommended\ Tags\}|} = \frac{|\{Dataset\ Tags\} \cap \{Recommended\ Tags\}|}{|\{Recommended\ Tags\}|} \quad 1)$$

Como segunda métrica se utilizó recall (2), que consiste en la fracción de etiquetas que son devueltas de forma exitosa.

$$Recall = \frac{Match}{Count} = \frac{|\{Dataset\ Tags\} \cap \{Recommended\ Tags\}|}{|\{Relevant\ Tags\}|} \quad 2)$$

Para realizar los cálculos se han utilizado las doce primeras etiquetas del vector de sugerencias, tanto para precisión@12 (léase precisión en 12) como recall@12 (léase recall en 12). Se ha seguido una propuesta similar a la de *Mishne et al.* [41], de igual forma, se asumió que para un usuario dado que no haya encuentra el grupo de etiquetas apropiada dentro de las primeras 12, este omitirá la búsqueda y añadirá una nueva de su puño y letra.

7.2 MUESTRA DE DATOS

Para realizar el experimento hemos utilizado una muestra de datos compuesta por 483 documentos a los que se han etiquetado un total de 1048 etiquetas. Este dataset ha sido provisto dentro del contexto del proyecto Origin desarrollado en colaboración con la empresa Treelogic S.L. Este conjunto de datos ha sido extraído de su herramienta de vigilancia estratégica de proyectos de investigación e I+D+I, etiquetado por los trabajadores de esta empresa en su trabajo diario. Este conjunto de documentos incluyen una proporción mayor de documentos en español que en inglés. Los diferentes documentos incluyen una diversidad de formatos (XML, HTML, texto plano) así como de longitud, incluyendo una elevada proporción de entradas "sucias" con pueden ser caracteres especiales XML o HTML.

Cada documento incluye al menos:

- Título de la Noticia
- Descripción
- Url relacionada
- ID del usuario
- Etiquetas separadas por punto y coma
- Fecha de creación

Los datos no ha sido pre-procesados o limpiados, buscando conseguir que las pruebas sean lo más fieles posible al comportamiento de ACOTA en un entorno real.

7.3 UNIDAD DE ANÁLISIS

Una vez definida la muestra del estudio es necesario definir qué características pueden y serán analizadas y la forma en que este análisis se llevará a cabo. Tomando como punto de partida los conceptos presentados, a continuación se muestras los elementos a ser medidos.

El análisis se realizará a todos los componentes de la metodología, es decir, se comprobará los criterios indicados en la sección 7.1, para los siguientes elementos:

- A.1 Normalización
- A.2 Procesamiento de Lenguaje Natural
- A.3 Agrupación
- B.1 Consultas a Ontologías
- B.2 Consultas a Folksonomías
- C.1 Recomendación de Etiquetas

Cabe resaltar que solo se han utilizado las funcionalidades integradas dentro de la librería ACOTA.

Capítulo 8. RESULTADOS OBTENIDOS

En esta sección se presenta los resultados de los experimentos llevados a cabo según las especificaciones descritas en la sección anterior. Los resultados de calcular precisión@12 y recall@12 para los distintas etapas de ACOTA están presentes en la Tabla 8.1, la Tabla 8.2 y la Tabla 8.3. Dichas tabas están divididas en tres columnas. En primer lugar precisión@12 está ubicada en la segunda columna mientras que recall@12 está ubicada en la tercera columna, respectivamente.

8.1 INTERCEPTACIÓN DE RESULTADOS

8.1.1A. Etapa de Extracción

Como era de esperar, la combinación de *A.1 Normalización* más el proceso de filtrado *A.2 procesamiento* de lenguaje natural con *A.3 Agrupamiento* obtuvo los máximos resultados dentro la etapa de extracción. Teniendo en cuenta que el conjunto de datos está compuesto tanto por unigramas como por n-gramas, la única forma de poder generar etiquetas que puedan validarse contra el conjunto completo, es mediante la combinación de las tres etapas. Es decir, en primer lugar mediante la extracción de unigramas (A1 + A2), y en segundo lugar mediante la de extracción de n-gramas (A3). Esto se debe, a que utilizando dichas técnicas por separado, solo sería posible validar unigramas o n-gramas en función de la técnica que se estuviese aplicando, (A1 + A2) o (A3) respectivamente.

Tabla 8.1 Etapa de Extracción, Precisión@12 y Recall@12

Objetivo	Precisión@12	Recall@12
A.1. Normalización	0.3737	0.0676
A.2. Normalización + NLP	0.3766	0.0682
A.3. Agrupamiento	0.1144	0.0207
A. Etapa de Extracción	**0.4109**	**0.0744**

8.1.2B. Etapa de Enriquecimiento

En el conjunto de datos predomina el español, disponiendo de un bajo porcentaje de documentos en otros idiomas (Inglés o Francés), por tanto, la etapa B.1 de consultas a ontologías no es tan efectiva como era de esperar (Esta etapa solo realiza recomendación de sinónimos para etiquetas en inglés). En contraste, las consultas a folksonomías incrementan los resultados tanto de precisión como de recall de forma notable, al ser este una etapa completamente multilingüe.

Tabla 8.2 Etapa de Enriquecimiento, Precisión@12 y Recall@12

Objetivo	Precisión@12	Recall@12
B.1. Consultas a Ontologías	0.4156	0.0752
B.2. Consulta a Folksonomías	0.4290	0.0776
B. Etapa de Enriquecimiento	**0.4290**	**0.0776**

8.1.3C. Etapa de Recomendación

Partiendo del sistema de persistencia vacío, según se iban recomendado documentos, los resultados exitosos, es decir aquellos que encajaban con los valores devueltos por los usuarios, se ha ido almacenando de vuelta al sistema (como feedback), tal y como habría se habría realizado por parte de los usuarios reales.

Este experimento se realizó siguiendo el mismo orden cronológico con el que se realizó el etiquetado de los documentos por parte de los usuarios, de esta forma se consigue reproducir las mismas condiciones con las que se realizaron la toma de muestras original. Como se puede observar existe una mejora apreciable tanto de precisión@12 como de recall@12, gracias a que el sistema tiene en cuenta el comportamiento previo de los usuarios con el sistema.

Tabla 8.3 Etapa de Recomendación, Precisión@12 y Recall@12

Objetivo	Precisión@12	Recall@12
C.1. Recomendación	0.4433	0.0802
C. Etapa de Recomendación	**0.4433**	**0.0802**

8.1.4Interpretación de los Resultados

Cabe mencionar que el conjunto de datos de entrada dispone de una gran cantidad de etiquetas que hacen referencia a eventos temporales, proyectos internos, o incluso asuntos personales. Por tanto, el número de etiquetas que son objetivo del sistema son menores que el número real de las mismas. Incluso el uso de una folksonomía como Google Complete API (Construida por miles de millones de personas en sus búsquedas diarias) no es capaz de recomendar dichas etiquetas.

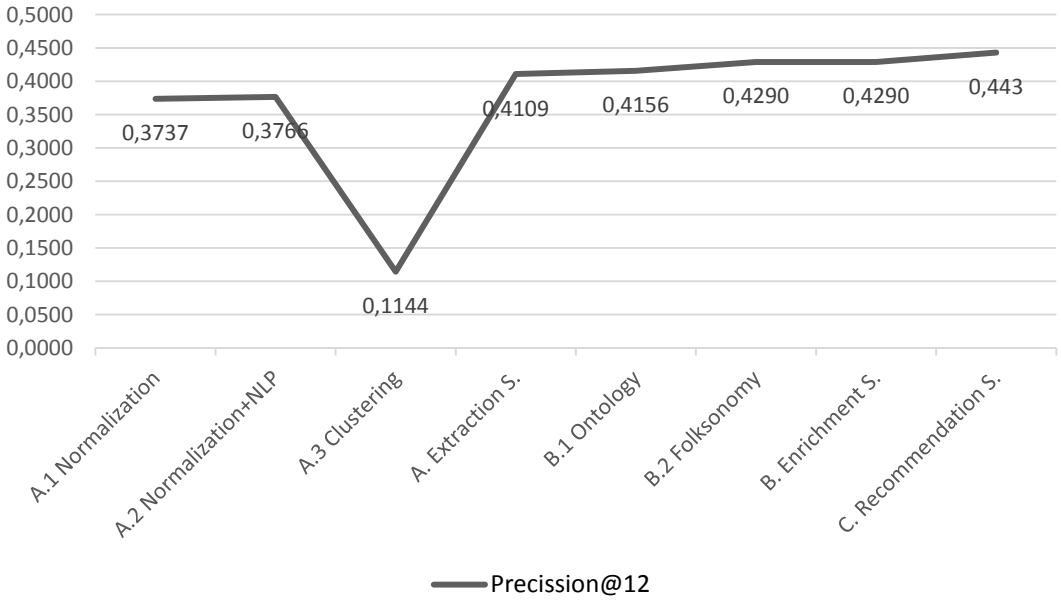


Figura 8.1 Precisión@12

Como puede ser apreciado en las diferentes tablas y gráficas, los valores que hacen referencia a recall@12 tienen unos valores considerablemente bajos. Teniendo en cuenta que el conjunto de datos provisto por Treelogic S.L contiene 1049 etiquetas para 483 documentos, la media de etiquetas por documento es de 2,1718 etiquetas/documento.

Partiendo de que para calcular la medida de recall@12 solo se tienen en cuenta las 12 primeras etiquetas dentro del vector de etiquetas, el resultado ideal de recall@12 sería de 0.1810 (Es decir 1049 etiquetas "objetivo" etiquetadas por los usuarios para 5749 etiquetas devueltas por ACOTA, a razón de 12 etiquetas por documento (483 documentos). Debido a estos motivos los resultados de recall@12 no están alejados como parecen, alcanzando valores relativos considerablemente buenos llegando a 42.90%.

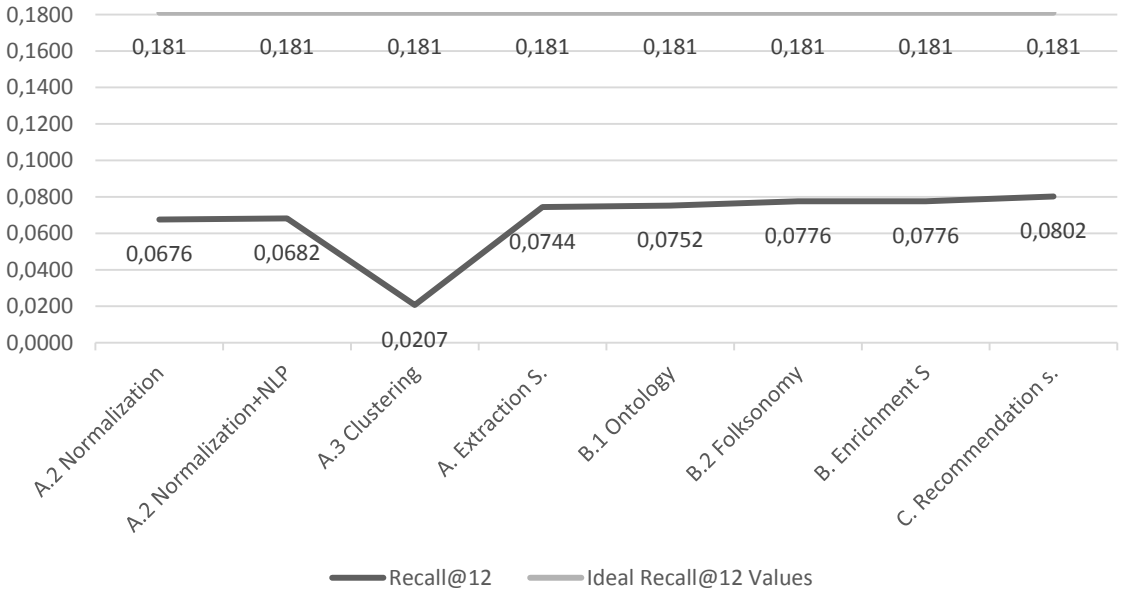


Figura 8.2 Recall@12

8.2 DISCUSIÓN

Los resultados obtenidos durante los experimentos llevados a cabo durando el presente proyecto, dan una idea del comportamiento de la metodología propuesta en un ambiente real, como lo es un sistema de alerta tecnológica. Mediante el desarrollo de estos experimentos se ha pretendido demostrar la efectividad real de la metodología propuesta que permite simplificar y favorecer el acto de etiquetación mediante la recomendación y sugerencia de las etiquetas.

Capítulo 9. CONCLUSIONES Y TRABAJO FUTURO

El presente proyecto, *ACOTA: TECNOLOGÍAS DE ETIQUETADO SEMIAUTOMÁTICO Y COLABORATIVO,* ha tenido como objetivo el desarrollo de una metodología multilingüe colaborativa y semiautomática. Dicha metodología tiene el cometido de reducir el esfuerzo asociado a la categorización de documentos, para ello nuestra propuesta utiliza técnicas de extracción de información, procesamiento de lenguaje natural y de aprendizaje automático. Esta metodología transfiere la decisión de elegir la etiqueta correcta al usuario final, de forma que aunque el sistema no recomiende una serie de etiquetas descriptivas, el usuario podría tomar la decisión de añadir nuevas etiquetas de su puño y letra, retroalimentando el sistema haciendo que esta aprenda mejorando los resultados de futuras recomendaciones.

La presente metodología ha tenido entre sus objetivos el de soportar la etiquetación de documentos multilingües, en gran medida debido al incremento de información en Internet disponible en idiomas distintos al inglés.

Además se ha tratado de que el sistema no sufra de arranque en frio, para ello se ha utilizado una aproximación hibrida, que combina un sistema de etiquetación automático con un sistema de recomendación de etiquetas. Esta hibridación permite que pese a que el sistema esté frio y el sistema de recomendación no pueda generar etiquetas válidas, el sistema de etiquetación automático sigue funcionando por debajo.

Como implementación de referencia se ha desarrollado una librería llamada ACOTA (*Automatic Collaborative Tagging*), implementa todos los pasos de la metodología propuesto. Ha sido testada contra una muestra de datos proveniente de un sistema de alerta tecnológica, en producción. A pesar de la cantidad de datos temporales, internos y de la cantidad de ruido como caracteres especiales, se ha conseguido alcanzar unos resultados aceptables. Esta metodología permite a los trabajadores del conocimiento (Knowledge Workers), disponer de información correctamente categorizada y disponible para su uso dentro de las diferentes organizaciones.

9.1 TRABAJO FUTURO

Pese a que todos los objetivos que se perseguían con este proyecto han sido completados de forma satisfactoria, quedan abiertas diferentes líneas de investigación y de desarrollo, que pueden ser utilizadas por otros investigadores o con el objetivo de mejorar los resultados de este proyecto:

- Aumentar el número de lenguajes nativos en el sistema, actualmente se da soporte completo a inglés y español. La inclusión de nuevos lenguajes permitirá mejorar los resultados y precisión del sistema con contenidos multilingües.
- Inclusión de compresión de etiquetas (tag compression). Esta técnica está altamente ligada a cada lenguaje, por lo que sería añadida ad-hoc en función de los requisitos. Permite agrupar contenidos sin importar el género y número del término, reduciendo

la cantidad de términos ambiguos dentro del sistema y por tanto, aumentando la precisión general del sistema.

- Aumentar el rendimiento del sistema. Si bien ACOTA tiene un buen rendimiento con ficheros regulares, sufre una penalización de rendimiento cuando se necesita etiquetar ficheros de gran tamaño.
- Implementación MapReduce[37] de la presente metodología. Si bien ACOTA no está pensado para etiquetar documentos de gran tamaño (BigData), si sería posible realizar pequeñas modificaciones para aprovechar las ventajas de ser ejecutado en un entorno distribuido como puede ser Apache Hadoop[43] o Spark[44].

9.2 DIFUSIÓN DE LOS RESULTADOS

9.2.1Trabajos Principales

Durante el presente proyecto, con el fin de difundir los resultados obtenidos por esta investigación se ha presentado un artículo a un congreso y otro a una revista dentro del índice de impacto JCR[45]. También se tiene previsto mandar un tercer artículo a un segundo congreso.

"*Towards a Multilingual - Domainless Automatic Collaborative Tagging Technique for Assisting Corporate Know-How Management*" ha sido presentado al *8th South East European Doctoral Student Conference* en Salónica (Grecia).

Estado: *Enviado, esperando revisión*.

"*Assisting Corporate Knowledge Management: A Multilingual and Domainless Semi-Automatic Collaborative Tagging Approach*", una versión refinada y completa del mismo ha sido presentado a la revista *Expert Systems with Applications (ESWA)* perteneciente a Elsevier. Dispone de un índice de impacto de 2,203 y de 2,455 a 5-años. *ESWA* es un referente internacional enfocado en sistemas expertos y sistemas inteligentes aplicados a diferentes campos como la industria, el gobierno y las universidades.

Estado: *Enviado, revisor asignado*.

Por último se tiene previsto mandar una versión modificada de los trabajos realizados, dando un mayor peso a gestión del conocimiento, al workshop *Knowledge and Experience Management* organizado por *Special Interest Group on Knowledge Management* de *la German Informatics Society* (GI).

Estado: *Trabajo Futuro*.

[43] http://hadoop.apache.org/
[44] http://spark-project.org/
[45] http://wokinfo.com/products_tools/analytical/jcr/

9.2.2Trabajos Suplementarios

Pese a haberse realizado con anterioridad al proyecto de fin de máster, como parte de una investigación previa, se ha colaborado en la publicación de un capítulo de un libro realizado por los miembros del grupo de investigación WESO (Web Semantics Oviedo).

Emergent Ontologies by collaborative tagging for Knowledge Management publicado como artículo en el libro *Advancing Information Management through Semantic Web Concepts and Ontologies*.

Estado: Publicado.

Capítulo 10. BIBLIOGRAFÍA

[1] N. Belkin and W. Croft, "Information filtering and information retrieval: two sides of the same coin?," *Communications of the ACM*, vol. 29, no. 10, pp. 1–10, 1992.

[2] "What is big data?," *IBM.com*, 2012. [Online]. Available: http://www-01.ibm.com/software/data/bigdata/. [Accessed: 08-May-2013].

[3] L. Shklar, A. Sheth, V. Kashyap, and K. Shah, "InfoHarness: Use of automatically generated metadata for search and retrieval of heterogeneous information," in *Advanced Information Systems Engineering*, 1995, pp. 217–230.

[4] A. Large and H. Moukdad, "Multilingual access to web resources: an overview," *Program: electronic library and information systems*, vol. 34, no. 1, pp. 43–58, 2000.

[5] W3Techs, "Usage of content languages for websites." [Online]. Available: http://w3techs.com/technologies/overview/content_language/all. [Accessed: 03-May-2013].

[6] M. de Kunder, "The size of the World Wide Web (The Internet)." [Online]. Available: http://www.worldwidewebsize.com/. [Accessed: 03-May-2013].

[7] B. Hjørland, "Semantics and knowledge organization," *Annual Review of Information Science and Technology*, vol. 41, no. 1, pp. 367–405, Oct. 2007.

[8] B. Hjørlandd, "Methods for evaluating information sources: An annotated catalogue," *Journal of Information Science*, vol. 38, no. 3, pp. 258–268, Apr. 2012.

[9] T. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?," *International Journal of Human-Computer Studies*, vol. 43, no. 5–6, pp. 907–928, Nov. 1995.

[10] T. Vander Wal, "Folksonomy Coinage and Definition," 2007. [Online]. Available: http://vanderwal.net/folksonomy.html.

[11] P. Levy, *Collective Intelligence: Mankind's Emerging World in Cyberspace*. Perseus Publishing, 1999.

[12] C. Casado-Lumbreras, A. Rodríguez-González, J. M. Álvarez-Rodríguez, and R. Colomo-Palacios, "PsyDis: Towards a diagnosis support system for psychological disorders," *Expert Systems with Applications*, vol. 39, no. 13, pp. 11391–11403, Oct. 2012.

[13] Á. García-Crespo, A. Rodríguez, M. Mencke, J. M. Gómez-Berbís, and R. Colomo-Palacios, "ODDIN: Ontology-driven differential diagnosis based on logical inference and probabilistic refinements," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2621–2628, Mar. 2010.

[14] B. Villazón-Terrazas, J. Ramírez, M. C. Suárez-Figueroa, and A. Gómez-Pérez, "A network of ontology networks for building e-employment advanced systems," *Expert Systems with Applications*, vol. 38, pp. 13612–13624, Apr. 2011.

[15] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Information retrieval in folksonomies: Search and ranking," *The Semantic Web: Research and Applications*, vol. 4011, pp. 411–426, 2006.

[16] D. Yoo, K. Choi, Y. Suh, and G. Kim, "Building and evaluating a collaboratively built structured folksonomy," *Journal of Information Science*, Mar. 2013.

[17] N. Tourné and D. Godoy, "Evaluating tag filtering techniques for web resource classification in folksonomies," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9723–9729, Aug. 2012.

[18] W. Carrer-Neto, M. L. Hernández-Alcaraz, R. Valencia-García, and F. García-Sánchez, "Social knowledge-based recommender system. Application to the movies domain," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10990–11000, Sep. 2012.

[19] C.-C. Kiu and E. Tsui, "TaxoFolk: A hybrid taxonomy–folksonomy structure for knowledge classification and navigation," *Expert Systems with Applications*, vol. 38, no. 5, pp. 6049–6058, May 2011.

[20] C. Shirky, "Ontology is Overrated: Categories, Links, and Tags," *Economics & Culture, Media & Community*, 2005. [Online]. Available: http://www.shirky.com/writings/ontology_overrated.html?goback=.gde_1838701_member_179729766.

[21] H. Halpin, V. Robu, and H. Shepard, "The dynamics and semantics of collaborative tagging," *Proceedings of the 1st Semantic Authoring and Annotation Workshop SAAW'06*, 2006.

[22] M. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, no. x, 2005.

[23] S.-T. Park, D. Pennock, O. Madani, N. Good, and D. DeCoste, "Naïve filterbots for robust cold-start recommendations," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, 2006, pp. 699–705.

[24] H. de Conceptos, "La Gestión del Conocimiento," *mineducacion.gov.co*, no. Gestión 2000, pp. 1–8, 2003.

[25] C. Koh and V. Prybutok, "The three ring model and development of an instrument for measuring dimensions of e-government functions," *Journal of Computer Information Systems*, vol. 43, no. 3, p. 6, 2003.

[26] C. S. Monreal and I. G. Leiva, "Posibilidades y límites de los tesauros frente a otros sistemas de organización del conocimiento :," pp. 361–377.

[27] S. Fernández, E. Rubiera, D. Berrueta, and L. Polo, "Estado del arte y propuesta de técnicas para la integración de ontologías y folksonomías," pp. 1–19, 2007.

[28] L. Peña, *Fundamentos de ontología dialéctica*. 1987.

[29] N. Guarino, "Formal Ontology in Information Systems," *Proceedings of FOIS'98*, no. June, pp. 3–15, 1998.

[30] P. Andersen, "What is Web 2.0?: ideas, technologies and implications for education," 2007.

[31] D. Berrueta-Muñoz, "MORFEO-MyMobileWebTecnologías avanzadas de software abierto para el desarrollo de la Web Móvil: 1.0, 2.0 y 3.0," 2006.

[32] M. Vallez and R. Pedraza-Jimenez, "El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines," *Hipertext.net*, vol. 5, 2007.

[33] S. Nakatani, "Language Detection Library for Java." p. 2012.

[34] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. 1999.

[35] A. Foundation, *OpenNLP Proposal*. .

[36] Mitchell, *Machine Learning*. McGraw Hill, 1997.

[37] J. Dean and S. Ghemawat, "MapReduce," *Communications of the ACM*, vol. 51, no. 1, p. 107, Jan. 2008.

[38] S. Owen, R. Anil, T. Dunning, and E. Friedman, *Mahout in action*. Shelter Island: Manning, 2011, p. 387.

[39] C. H. Brook and N. Montanez, "Improved annotation of the blogopshere via autotagging and hierarchical clustering," *Proceedings of the 15th World Wide Web Conference (WWW06)*, 2006.

[40] P. Soucy and G. W. Mineau, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model."

[41] G. Mishne, "AutoTag," in *Proceedings of the 15th international conference on World Wide Web (WWW 06)*, 2006, p. 953.

[42] S. C. Sood, S. H. Owsley, K. J. Hammond, and L. Birnbaum, "TagAssist: Automatic Tag Suggestion for Blog Posts," 2007.

[43] K. Sun, X. Wang, C. Sun, and L. Lin, "A language model approach for tag recommendation," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1575–1582, Mar. 2011.

[44] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles, "Real-time automatic tag recommendation," *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, p. 515, 2008.

[45] P.-I. Chen and S.-J. Lin, "Automatic keyword prediction using Google similarity distance," *Expert Systems with Applications*, vol. 37, no. 3, pp. 1928–1938, Mar. 2010.

[46] W. Jimenez-Nácero, C. Luis-Alvargonzález, P. Abella-Vallina, J. M. Alvarez-Rodríguez, J. E. Labra-Gayo, and P. Ordoñez de Pablos, "Emergent Ontologies by collaborative tagging for Knowledge Management," in *Advancing Information Management through Semantic Web Concepts and Ontologies*, IGI-Global, 2012, p. 16.

[47] B. Sigurd, M. Eeg-Olofsson, and J. van Weijer, "Word length, sentence length and frequency - Zipf revisited," *Studia Linguistica*, vol. 58, no. 1, pp. 37–52, Apr. 2004.

[48] F. M. Suchanek, M. Vojnovic, and D. Gunawardena, "Social tags," in *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, 2008, p. 223.

[49] M. Dostal and K. Ježek, "Automatic keyphrase Extraction extraction based on NLP Automatic Keyphrase based on NLP and statistical methods and Statistical Methods," *Proceedings of the Dateso 2011: Annual International Workshop on DAtabases, TExts, Specifications and Object*, pp. 140–145, 2011.

[50] G. a. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[51] R. T. Fielding, "Architectural Styles and the Design of Network-based Software Architectures," Irvine: , 2000.

[52] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the tenth international conference on World Wide Web - WWW '01*, 2001, pp. 285–295.

[53] T. T. Tanimoto, "An elementary mathematical theory of classification and prediction," 1958.

[54] A. Lipkus, "A proof of the triangle inequality for the Tanimoto distance," *Journal of Mathematical Chemistry*, vol. 26, pp. 263–265, 1999.

[55] C. H. Davidson, "Technology watch in the construction sector: why and how?," *Building Research & Information*, vol. 29, no. 3, pp. 233–241, May 2001.

[56] J. Keats, *Virtual Words: Language on the Edge of Science and Technology*. Oxford, New York: Oxford University Press, Inc., 2010.

[57] C. Cleverdon, J. Mills, and M. Keen, "Factors determining the performance of indexing systems," 1966.

Capítulo 11. APÉNDICES

11.1 LICENCIAS

A continuación se muestran las licencias bajo las que ha sido licenciado el código desarrollado en este proyecto.

11.1.1 Apache License 2.0

Todo el código fuente desarrollado en el presente proyecto ha sido licenciado bajo licencia Apache 2.0[46].

```
                                 Apache License
                           Version 2.0, January 2004
                        http://www.apache.org/licenses/

   TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION

   1. Definitions.

      "License" shall mean the terms and conditions for use, reproduction,
      and distribution as defined by Sections 1 through 9 of this document.

      "Licensor" shall mean the copyright owner or entity authorized by
      the copyright owner that is granting the License.

      "Legal Entity" shall mean the union of the acting entity and all
      other entities that control, are controlled by, or are under common
      control with that entity. For the purposes of this definition,
      "control" means (i) the power, direct or indirect, to cause the
      direction or management of such entity, whether by contract or
      otherwise, or (ii) ownership of fifty percent (50%) or more of the
      outstanding shares, or (iii) beneficial ownership of such entity.

      "You" (or "Your") shall mean an individual or Legal Entity
      exercising permissions granted by this License.

      "Source" form shall mean the preferred form for making modifications,
      including but not limited to software source code, documentation
```

[46] http://www.apache.org/licenses/LICENSE-2.0.txt

```
           source, and configuration files.

           "Object" form shall mean any form resulting from mechanical
           transformation or translation of a Source form, including but
           not limited to compiled object code, generated documentation,
           and conversions to other media types.

           "Work" shall mean the work of authorship, whether in Source
or
           Object form, made available under the License, as indicated
by a
           copyright notice that is included in or attached to the work
           (an example is provided in the Appendix below).

           "Derivative Works" shall mean any work, whether in Source or
Object
           form, that is based on (or derived from) the Work and for
which the
           editorial revisions, annotations, elaborations, or other
modifications
           represent, as a whole, an original work of authorship. For
the purposes
           of this License, Derivative Works shall not include works
that remain
           separable from, or merely link (or bind by name) to the
interfaces of,
           the Work and Derivative Works thereof.

           "Contribution" shall mean any work of authorship, including
           the original version of the Work and any modifications or
additions
           to that Work or Derivative Works thereof, that is
intentionally
           submitted to Licensor for inclusion in the Work by the
copyright owner
           or by an individual or Legal Entity authorized to submit on
behalf of
           the copyright owner. For the purposes of this definition,
"submitted"
           means any form of electronic, verbal, or written
communication sent
           to the Licensor or its representatives, including but not
limited to
           communication on electronic mailing lists, source code
control systems,
           and issue tracking systems that are managed by, or on behalf
of, the
           Licensor for the purpose of discussing and improving the
Work, but
           excluding communication that is conspicuously marked or
otherwise
           designated in writing by the copyright owner as "Not a
Contribution."

           "Contributor" shall mean Licensor and any individual or Legal
Entity
           on behalf of whom a Contribution has been received by Licensor
and
           subsequently incorporated within the Work.

        2. Grant of Copyright License. Subject to the terms and
conditions of
```

```
            this License, each Contributor hereby grants to You a
perpetual,
            worldwide, non-exclusive, no-charge, royalty-free,
irrevocable
            copyright license to reproduce, prepare Derivative Works of,
            publicly display, publicly perform, sublicense, and
distribute the
            Work and such Derivative Works in Source or Object form.

        3. Grant of Patent License. Subject to the terms and conditions
of
            this License, each Contributor hereby grants to You a
perpetual,
            worldwide, non-exclusive, no-charge, royalty-free,
irrevocable
            (except as stated in this section) patent license to make,
have made,
            use, offer to sell, sell, import, and otherwise transfer the
Work,
            where such license applies only to those patent claims
licensable
            by such Contributor that are necessarily infringed by their
            Contribution(s) alone or by combination of their
Contribution(s)
            with the Work to which such Contribution(s) was submitted.
If You
            institute patent litigation against any entity (including a
            cross-claim or counterclaim in a lawsuit) alleging that the
Work
            or a Contribution incorporated within the Work constitutes
direct
            or contributory patent infringement, then any patent licenses
            granted to You under this License for that Work shall
terminate
            as of the date such litigation is filed.

        4. Redistribution. You may reproduce and distribute copies of
the
            Work or Derivative Works thereof in any medium, with or
without
            modifications, and in Source or Object form, provided that
You
            meet the following conditions:

            (a) You must give any other recipients of the Work or
                Derivative Works a copy of this License; and

            (b) You must cause any modified files to carry prominent
notices
                stating that You changed the files; and

            (c) You must retain, in the Source form of any Derivative
Works
                that You distribute, all copyright, patent, trademark,
and
                attribution notices from the Source form of the Work,
                excluding those notices that do not pertain to any part
of
                the Derivative Works; and

            (d) If the Work includes a "NOTICE" text file as part of its
```

distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.

You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

5. Submission of Contributions. Unless You explicitly state otherwise, any Contribution intentionally submitted for inclusion in the Work by You to the Licensor shall be under the terms and conditions of this License, without any additional terms or conditions. Notwithstanding the above, nothing herein shall supersede or modify the terms of any separate license agreement you may have executed with Licensor regarding such Contributions.

6. Trademarks. This License does not grant permission to use the trade names, trademarks, service marks, or product names of the Licensor, except as required for reasonable and customary use in describing the origin of the Work and reproducing the content of the NOTICE file.

```
        7. Disclaimer of Warranty. Unless required by applicable law or
           agreed to in writing, Licensor provides the Work (and each
           Contributor provides its Contributions) on an "AS IS" BASIS,
           WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express
or
           implied, including, without limitation, any warranties or
conditions
           of TITLE, NON-INFRINGEMENT, MERCHANTABILITY, or FITNESS FOR
A
           PARTICULAR PURPOSE. You are solely responsible for
determining the
           appropriateness of using or redistributing the Work and
assume any
           risks associated with Your exercise of permissions under this
License.

        8. Limitation of Liability. In no event and under no legal
theory,
           whether in tort (including negligence), contract, or
otherwise,
           unless required by applicable law (such as deliberate and
grossly
           negligent acts) or agreed to in writing, shall any Contributor
be
           liable to You for damages, including any direct, indirect,
special,
           incidental, or consequential damages of any character arising
as a
           result of this License or out of the use or inability to use
the
           Work (including but not limited to damages for loss of
goodwill,
           work stoppage, computer failure or malfunction, or any and
all
           other commercial damages or losses), even if such Contributor
           has been advised of the possibility of such damages.

        9. Accepting Warranty or Additional Liability. While
redistributing
           the Work or Derivative Works thereof, You may choose to offer,
           and charge a fee for, acceptance of support, warranty,
indemnity,
           or other liability obligations and/or rights consistent with
this
           License. However, in accepting such obligations, You may act
only
           on Your own behalf and on Your sole responsibility, not on
behalf
           of any other Contributor, and only if You agree to indemnify,
           defend, and hold each Contributor harmless for any liability
           incurred by, or claims asserted against, such Contributor by
reason
           of your accepting any such warranty or additional liability.

        END OF TERMS AND CONDITIONS

        APPENDIX: How to apply the Apache License to your work.

           To apply the Apache License to your work, attach the following
           boilerplate notice, with the fields enclosed by brackets "[]"
```

```
           replaced with your own identifying information. (Don't
include
           the brackets!) The text should be enclosed in the appropriate
           comment syntax for the file format. We also recommend that a
           file or class name and description of purpose be included on
the
           same "printed page" as the copyright notice for easier
           identification within third-party archives.

        Copyright [yyyy] [name of copyright owner]

        Licensed under the Apache License, Version 2.0 (the "License");
        you may not use this file except in compliance with the License.
        You may obtain a copy of the License at

            http://www.apache.org/licenses/LICENSE-2.0

        Unless required by applicable law or agreed to in writing,
software
        distributed under the License is distributed on an "AS IS" BASIS,
        WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or
implied.
        See the License for the specific language governing permissions
and
        limitations under the License.
```

11.2 APACHE 2.0 LICENSE - CASTELLANO

Esta tradición no es una traducción oficial de la licencia Apache License 2.0, está realizada únicamente como traducción orientativa y por tanto no obliga a cumplir los términos de la misma.

```
                                 Apache License
                           Version 2.0, January 2004
                        http://www.apache.org/licenses/

     CONDICIONES DE USO, REPRODUCCIÓN Y DISTRIBUCIÓN

     1. Definiciones.

     "Licencia" hace referencia a las condiciones de uso, reproducción
y distribución, según la definición establecida en las Secciones 1 a 9
de este documento.

     "Licenciador" hace referencia al propietario de los derechos de
autor, o la entidad autorizada por el mismo, que otorga la licencia.

     "Entidad legal" hace referencia a la unión de la entidad actuante
y todas las demás entidades que la controlan, son controladas por ella
o están sujetas a un control común con dicha entidad. Para los fines de
esta definición, el "control" es (i) la potestad directa o indirecta
para dirigir dicha entidad, ya sea mediante contrato o de otro modo; o
(ii) la titularidad de al menos un cincuenta por ciento (50 %) de las
acciones, o (iii) la propiedad efectiva de dicha entidad.

     "Usted" (o "Su") hace referencia a una persona o entidad legal que
ejerza las autorizaciones otorgadas por esta Licencia.
```

La forma del "Código" hace referencia a la forma preferente de realizar modificaciones, como por ejemplo, el código fuente del software, la fuente de la documentación o los archivos de configuración.

La forma del "Objeto" hace referencia a cualquier forma resultante de la transformación mecánica o traducción de una forma de Código, como por ejemplo, código de objeto compilado, documentación generada y conversiones a otros tipos de medios.

"Obra" hace referencia a la obra de autor, ya sea en forma de Código o de Objeto, disponible en virtud de esta Licencia, según la indicación de copyright incluida o incorporada a la obra (se facilita un ejemplo en el apéndice más adelante).

"Obras derivativas" hace referencia a todas las obras, tanto en forma de Código como de Objeto, que estén basadas en (o derivadas de) la Obra y en las que el conjunto de las revisiones de modificación, anotaciones, elaboración y otros cambios representan, en total, una obra de autoría original. Para los fines de esta Licencia, las Obras Derivativas no incluirán las obras que sea posible separar o que compartan un simple vínculo (o unión por nombre) con las interfaces de la Obra y de las Obras Derivativas de la misma.

"Contribución" hace referencia a cualquier obra de autor, incluida la versión original de la Obra y todas las modificaciones y adiciones a dicha Obra u Obras Derivativas de la misma, que se envíen deliberadamente al Licenciador para su inclusión en la Obra por el titular de los derechos de autor o por una persona o entidad legal autorizada para ello en representación del titular de los derechos. Para los fines de esta definición, enviar hace referencia a cualquier forma de comunicación electrónica, verbal o escrita efectuada por el Licenciador o sus representantes, como por ejemplo, la comunicación en listas de correo electrónico, sistemas de control de código fuente y sistemas de seguimiento de problemas gestionados por o en representación del Licenciador con el fin de comentar y mejorar la Obra, pero con exclusión de las comunicaciones descritas claramente o designadas por escrito por el titular de los derechos como "No contribución."

"Contribuyente" hace referencia al Licenciador o cualquier persona o Entidad Legal en cuyo nombre haya recibido el Licenciador una Contribución que se incorpore posteriormente a la Obra.

2. Concesión de licencia de derechos de autor. De acuerdo con las condiciones de esta Licencia, por la presente, cada Contribuyente le otorga a Usted una licencia de derechos de autor irrevocable, perpetua, internacional, no exclusiva, sin cargas ni regalías, para reproducir, preparar Obras Derivativas, mostrar u operar públicamente, sublicenciar y distribuir la Obra y tales Obras Derivativas en forma de Código o de Objeto.

3. Concesión de licencia patente. De acuerdo con las condiciones de esta Licencia, por la presente, cada Contribuyente le otorga a Usted una licencia de patente irrevocable (excepto en los casos indicados en esta sección), perpetua, internacional, no exclusiva, sin cargas ni regalías, para utilizar o hacer utilizar, vender, ofertar, importar y transferir de otro modo la Obra, siendo de aplicación esta Licencia

solamente a las solicitudes de patente otorgables por un Contribuyente que afecten exclusivamente a su(s) Contribución(es) o combinación de sus Contribuciones a la Obra a la que se incorporaron. Si usted iniciara un proceso legal contra cualquier entidad (incluidas las contrademandas en un proceso) alegando que la Obra o una Contribución incorporada a la Obra constituye una infracción directa o contributiva de una patente, todas las licencias de patentes otorgadas a Usted en virtud de esta Licencia sobre esa Obra cesarán a partir de la fecha en que se instituya dicho proceso.

4. Redistribución. Usted podrá reproducir y distribuir copias de la Obra o de las Obras Derivativas de la misma en cualquier medio, con o sin modificaciones, en forma de Código o de Objeto, siempre que cumpla las siguientes condiciones:

Deberá facilitar a todos los demás receptores de la Obra u Obras Derivativas una copia de esta Licencia y

Deberá indicar claramente las modificaciones que haya realizado en los archivos; y

Deberá conservar en la forma de Código de todas las Obras Derivativas que Usted distribuya todas las advertencias relativas a los derechos de autor, patentes, marcas registradas y atribuciones en la forma de Código de la Obra, con exclusión de las advertencias que no pertenezcan a ninguna parte de las Obras Derivativas, y

Si la Obra incluye un archivo de texto de "ADVERTENCIA/AVISO" como parte de su distribución, todas las Obras Derivativas que Usted distribuya deberán incluir una copia legible de las advertencias de atribución contenidas en dicho archivo de ADVERTENCIA, con exclusión de las advertencias que no pertenezcan a ninguna parte de las Obras Derivativas, en al menos uno de los lugares siguientes: En el contenido del archivo de texto de ADVERTENCIA o AVISO distribuido como parte de la Obra Derivativa; en el contenido de la forma de Código o en la documentación, si estos se facilitan junto con las Obras Derivativas; o mediante un dispositivo de muestra generado por las Obras Derivativas, en el lugar en que normalmente aparezcan tales advertencias externas. El contenido del archivo de ADVERTENCIA es meramente informativo y no modifica la Licencia. Usted podrá añadir Sus propias advertencias de atribución en las Obras Derivativas que Usted distribuya, junto con el texto de ADVERTENCIA de la Obra, o como adición al mismo, siempre que estas advertencias de atribución adicionales no puedan interpretarse como una modificación de la Licencia. Usted podrá añadir Su propia declaración respecto a los derechos de autor en Sus modificaciones y podrá añadir distintas condiciones en la Licencia para el uso, la reproducción o la distribución de Sus modificaciones o de las Obras Derivativas en conjunto, siempre que Su uso, reproducción y distribución de la Obra cumpla por lo demás las condiciones establecidas en esta Licencia.

5. Envío de contribuciones. A menos que Usted indique lo contrario, todas las Contribuciones que Usted envíe deliberadamente al Licenciador para su inclusión en la Obra, estarán sujetas a las condiciones de esta Licencia sin aplicación de condiciones adicionales. No obstante, ninguna disposición de este documento invalidará ni modificará las condiciones de cualquier otro acuerdo de Licencia que usted pueda haber suscrito con el Licenciador en relación con tales Contribuciones.

6. Marcas comerciales. Esta Licencia no autoriza el uso de nombres comerciales, marcas comerciales, marcas de servicios o nombres de productos del Licenciador, excepto cuando lo requiera el uso razonable y habitual en la descripción del origen de la Obra y la reproducción del contenido del archivo de ADVERTENCIA.

7. Exención de garantía. A menos que lo exijan las leyes pertinentes o se acuerde por escrito, el Licenciador ofrece la Obra (y cada Contribuyente ofrece sus Contribuciones) "TAL CUAL", SIN GARANTÍAS NI CONDICIONES DE NINGÚN TIPO, ya sean expresas o implícitas, como por ejemplo, cualquier garantía o condición sobre TÍTULO, NO INFRACCIÓN, APTITUD PARA EL COMERCIO o IDONEIDAD PARA UN FIN PARTICULAR. Usted es el/la único(a) responsable de determinar si es apropiado utilizar o redistribuir la Obra y asume todos los riesgos asociados a Su ejercicio de los permisos otorgados en esta Licencia.

8. Responsabilidad limitada. Bajo ninguna circunstancia ni fundamento legal, sea por ilícito civil extracontractual (incluida la negligencia), por contrato o de otro modo, a menos que lo exijan las leyes pertinentes (como en el caso de actos de negligencia deliberados y graves) o se haya acordado por escrito, será responsable ningún Contribuyente ante Usted por daños de ningún tipo, ya sean directos, indirectos, especiales, incidentales o consecuentes ocasionados como resultado de esta Licencia o por el uso o imposibilidad de uso de la Obra (como por ejemplo los daños por pérdida de clientes, pérdida de actividad, avería o mal funcionamiento de los ordenadores o cualquier otra forma de perjuicios o pérdidas comerciales), incluso si dicho Contribuyente hubiese sido advertido de la posibilidad de tales perjuicios.

9. Aceptación de garantías o responsabilidad adicional. En Su redistribución de la Obra o de las Obras Derivativas de la misma, Usted podrá ofrecer y cobrar por la aceptación de asistencia, garantías, indemnización u otras obligaciones de responsabilidad y/o derechos en virtud de esta Licencia. No obstante, al aceptar tales obligaciones, Usted podrá actuar solamente en Su propio nombre y bajo Su propia responsabilidad, no en nombre de ningún otro Contribuyente, y solamente si Usted accede a indemnizar, defender y eximir a cada Contribuyente de cualquier tipo de responsabilidad o disputas contra dicho Contribuyente, como resultado de u aceptación de tales garantías o responsabilidades adicionales.

FIN DE LAS CONDICIONES

APÉNDICE: Cómo aplicar la Licencia de Apache a su obra

Para aplicar la Licencia de Apache a su obra, añada el siguiente aviso, reemplazando el espacio en el interior de los corchetes "[]" con su propia información identificativa (¡no incluya los corchetes!). El texto deberá adaptarse a la sintaxis del texto correspondiente en el archivo. También recomendamos la inclusión de un nombre de archivo o de clase y una descripción de finalidad en la misma página impresa que la advertencia sobre derechos de autor, para facilitar la identificación en los archivos de terceros.

Derechos de autor [yyyy] [nombre del propietario de los derechos de autor]

Capítulo 12. PUBLICACIONES ADJUNTAS

1. Towards a Multilingual - Domainless Automatic Collaborative Tagging Technique for Assisting Corporate Know-How Management
2. Assisting Corporate Knowledge Management: A Multilingual and Domainless Semi-Automatic Collaborative Tagging Approach

TOWARDS A MULTILINGUAL - DOMAINLESS AUTOMATIC COLLABORATIVE TAGGING TECHNIQUE FOR ASSISTING CORPORATE KNOW-HOW MANAGEMENT

César Luis-Alvargonzález[1], Jose María Alvarez-Rodríguez[2] and Jose Emilio-Labra Gayo[1],

[1] WESO Research Group, Department of Computer Science, University of Oviedo, 33005, Oviedo, Spain
cesar.luis@weso.es, labra@uniovi.es
[2]South East European Research Center, 54622, Thessaloniki, Greece
jmalvarez@seerc.org

Abstract. The present paper introduces a hybrid methodology to automatically deploy and combine collaborative tagging techniques based on user-behaviour and well-known recommendation algorithms to name a few. A reference implementation called ACOTA (Automatic Collaborative Tagging) is also outlined in order to demonstrate the new recommendation capabilities of this approach with the aim of assisting both expert and non-expert users when multilingual resource tagging is required. Finally a quantitative research study in the context of corporate know-how management is also presented to evaluate and assess the goodness and accuracy of the methodology to minimize the effort of multilingual document categorization.

Keywords: Tagging, Collaborative Tagging, Automatic Tagging, Multilingual data, Real-Time, Information Retrieval, Document Retrieval

1 Introduction

The sheer mass of data already available in the Internet and the increasing use of mobile devices such as smart-phones, tablets or e-books readers are generating a new and dynamic data/information realm in which new requirements are emerging [1]. Some time ago information resources were printed on paper and kept in cabinets but nowadays this new vast amount of data [2] (in 2012, 2.5 exabytes of information were produced each day on the Internet) is commonly stored in digital formats with different organization and access techniques with the objective of easing processes such as information and document retrieval or search and report, to name a few. Although they try to take advantage of indexing and categorizing algorithms, on-line information resources usually contain a lot of heterogeneities [3] that cannot be easily addressed. More specifically, the multilingual encoding of information is one of the main drawbacks to provide a common access to the information [4]. In this sense, Internet

contents are currently available on a huge variety of languages but English is the most common language [5], used in 54.9% of websites, and most of the tools and techniques to perform some kind of exploitation over this information are customized to work on it. This implies that the rest of information (45.1%) faces a lack of tools to properly exploit this information. Taking into account that there are around 14.24 billion indexed web sites [6], 6.44 billion of web sites are in other languages rather than English, and as a consequence new tools are required to bridge the gap between the exploitation capabilities of English and non-English information resources.

Furthermore organizing this vast amount of multilingual data can be tedious, but in some enterprise and academics fields, it can become into the cornerstone to success. E-learning systems, B2B environments, Corporate Knowledge Management or Extraction and Information Retrieval are some domains in which data filtering and management processes are becoming crucial to get more accurate and timely information with the aim of enabling new business opportunities. More specifically knowledge organizations [7] and workers have gained their momentum in the new information society. Activities such as efficient document description, indexing or classification of information resources are now major challenges due to the aforementioned dynamic data realm. In this sense Hjørland [8] establishes different approaches such as use of controlled domain-based vocabularies and information retrieval techniques, user-oriented cognitive views or bibliometric approaches among others in order to ease the tasks of knowledge workers. As a consequence organizations with the capability of managing their knowledge activities can take advantage of discovering new business opportunities or workers capabilities due to the analysis of their daily activities. In this sense technology forecasting is one of the key activities in some of these companies (as an example Treelogic S.L.[1], a technology company, has established a percentage of the working time for this task) in which workers add everyday news, blog posts, research articles, funding opportunities, etc. that help companies to be aware of the new trends. In this environment the proper classification and exploitation of information resources leads us to a knowledge organization that can exploit this information and, as a consequence, its know-how for improving their own performance.

In order to address these new requirements in knowledge organizations there are several well-known and traditional techniques to model, structure, organise and exploit data and information such as conceptual maps, taxonomies or controlled vocabularies. Nevertheless recent times have shown a growing use of ontologies and folksonomies as a method to efficiently manage a vast amount of information resources. In this particular case, an ontology can be defined as a specification of a shared conceptualization [9], it is a formal description of concepts and their relationships involved in knowledge domain that assist to organize and build knowledge-based system. On the other hand, a folksonomy [10] is the result of free categorization, without a defined structure or formality. Usually it is created by collective intelligence; a group of users with different levels of knowledge interested in some domain collect and classify information resources. These two emerging approaches have been successfully applied to particular domains such as e-Health, e-Government or e-Procurement (ontologies [11]) and Web 2.0 sites (folksonomies [12][13]) such as Delicious, Flickr, Youtube, Linkedin, Twitter or Facebook.

[1] http://www.treelogic.com

In general, ontologies are more adequate to classify domain-specific information and data in a restricted context in which experts in a field have reached a common and shared understanding. In this context, data, information and knowledge is commonly concrete, static and stable. Obviously this approach presents some drawback to tackle data, information and knowledge management in a more general and dynamic environment. Nevertheless some structural upper ontologies such as DOLCE, SUMO or PROTON have been delivered in order to formalize high-level entities and relationships but their adoption to solve practical problems is still missing. On the other hand, as Shirky pointed in [14] a folksonomy works better with a large and dynamic corpus, unstable entities and participants of all levels of knowledge that seem to fit better to the existing web realm trying to exploit information to provide business intelligent services such as marketing suggestions or knowledge discovery. Furthermore another issue to take into account in data and information management lies in user behaviour, some works [15], have demonstrated that once a tag has been included within the system, there is a high probability that it will be used it again, obeying the Pareto principle rule 80-20 [16].

According to the aforementioned key points: a growing and dynamic new data realm, necessities in knowledge organizations and use of ontologies and folksonomies as knowledge management approaches, a hybrid methodology, employing automatic tagging techniques and user-behaviour recommendation algorithms that take advantage of folksonomies as previous step to consolidate knowledge in an ontology is presented. This methodology is built on the top of different techniques as natural language processing, querying to both folksonomies and ontologies, collaborative or machine learning techniques among others to deliver a multilingual methodology which does not suffer from a cold start [17]. The Automatic COllaborative Tagging (hereafter ACOTA) java-based library is the reference implementation that provides an automatic tagging engine with collaborative and user-behaviour recommendation capabilities. Finally, a quantitative evaluation taking into account existing information retrieval measures such as precision and recall is also performed over a real and private dataset (480 tagged information resources) created by the employees of Treelogic S.L. with the objective of assessing the tagging capabilities of the ACOTA library.

The rest of the paper is structured as follows. Section 2 describes the related work. Section 3 presents the proposed methodology. An overview of the architecture to support a methodology for multilingual and domain-less collaborative tagging is in the section 4. Sections 5 presents respectively the experiment methodology, evaluation, and the discussion and finally in the section 6 we face the conclusions and future work.

2 Related Work

The literature review of this work covers existing techniques and approaches to manage information resources in a knowledge environment. Thus y taking into account the use of blogs as information sources there are a certain amount of projects [18][19][20] trying to generate tags from these documents. Although information knowledge goes beyond blog-post tagging, it has a certain similarity, since a blog-post can be seen as traditional document, that cover many and specific different topics,

rather than web sites. In this sense, Brooks et al. do research [18] built a system that was based on the use of the top three term frequency–inverse document frequency (TFIDF) score tags of the blog post. This approach groups documents into clusters with the hypothesis that a cluster of documents that shared a tag should be more similar than a randomly constructed set of documents, the proposes methodology uses a similar approach but with a different purpose, to suggest for a given tag, common tags that are within the same cluster. AutoTag developed by Mishne et al. [19] is a tag suggestion blog post engine which employs collaborative filtering methods. This tool offers suggestions for tags based on tags assigned to similar posts, leaving the editor the decision of choosing the proper set of tags. In the particular case of the presented work, the presented methodology also transfers the final decision to the user, and employs similar suggestion techniques, but it does not rely only on collaborative methods, adding a pre-automatic tag generation step, based on the structure of the document. Finally, TagAssist is another automatic tag suggestion engine developed by Sood et al. [20], it evolves AutoTag design, adding support for tag compression. This approach provides a good support for English posts that serves as inspiration for the proposed multilingual methodology but it is exclusive for this language, in contrast with our solution. Furthermore and due to the growing use of social networks, some works have emerged to detect experts [21], trends and filter information [22] among others. The approach here presented must take advantage of these existing works to adapt broad techniques in a narrower context such as know-how management in a company.

In the case of Automatic Keywords Extraction [23][24], Dostal et al. [25] developed an automatic keyphrase extraction based on Natural Language Processing (hereafter NLP) and statistical methods, it applies to the presented paper due to the approach also includes NLP techniques, but employing a more customized and domain-based NLP techniques than this previous research. Although these existing works are focused on English tag recommendation, the main drawback lies in their lack of being configured to support new languages. As previous section has introduced this is not a mere issue, taking into account that English is just the third native language most spoken in the world [26] and a lot of digital information is written in other languages. In the case of TagAssist [20] and Brooks [18] they drop out any non-English words; on the other hand AutoTag [19] takes a less aggressive solution, holding non-English words, but due to the application of an English stemmer for any word, regardless the language, it ends up giving low scores to non-English posts. In the case of ACOTA, as a multilingual automatic and collaborative tagging engine, it is focused on providing support for non-English languages, e.g. Spanish due to the motivating scenario, in addition to English. That is why an initial design requirement lies in easing the configuration and the extension for other languages.

Most of the aforementioned works suffer from the well-known issue in recommendation systems, the cold start [17]. It appears when the system is relative new or unused, so there is no enough data to properly perform the recommendation algorithms. This issue is partially addressed by the methodology and ACOTA and it is able to suggest comprehensive tags even when there is no data stored on the system due to part of the suggestion takes as input public existing folksonomies. In this sense, a similar approach has been followed in previous works [27][28] that have been made use of queries to folksonomies such as Google Complete API to enrich tags suggestion. Nevertheless in the context of the present paper results of empirical experiments show

a considerable amount of noise that decreases the accuracy of suggestions. For instance the delegation of suggestions to this service without any pre/post processing implies the generation of a vector of suggestions with n tags, where some tags are representative according to the initial query but they are not usually representative enough for a domain-based recommendation. In order to partially reuse this service and avoid non-representative tags ACOTA prepares the queries and filters the external suggestions according to the existing domain.

3 A Multilingual - Domainless Automatic Collaborative Tagging Methodology

In this section a detailed description of the proposed methodology is provided. The methodology and, as a consequence ACOTA, is divided into two main stages: 1) the automatic tagging engine that is in charge of normalizing text, extract keywords, etc. and 2) the recommendation tagging engine that uses the outcomes of the first stage to finally generate suggest tags in a certain context.

3.1 Automatic Tagging Engine

The automatic tagging engine is comprised of two sequential stages. Firstly, the extraction stage which consists on extracting tags based on the document structure is applied to an existing dataset. Afterwards, the previous results are enhanced in the enrichment stage, querying to both folksonomies and ontologies. As a result, a set of descriptive tags are generated by the automatic tagging engine. Previously to this process and as usual in any technique dealing with natural language, an optional but relevant step must be executed in order to clean the raw text, removing special characters, stop-words, etc. The aim of this cleaning is to decrease the potential noise and to avoid the spreading of these non-useful lemmas to further stages.

A. Extraction Stage. This stage extracts tags from a document. It is composed of two different extracting techniques: 1) unigrams generation, it is performed by normalization and NLP, and 2) n-grams generation, a refined combination of the aforementioned techniques.

A.1. Normalization. This phase takes as input an information resource and produces a vector of unigrams. The first step is to split words at punctuation characters, removing punctuation marks and other delimiters. Then, each token is turned to a lower case. Tokens with only one character are removed. Finally, tokens are filtered according to a stop-word set (depending on the language common stop-words sets are available and can be configured), ending up with a clean set of unigram tags.

For each tag the number of occurrences is counted, multiplying this number by a weight. According to the word position or appearance, the weight is calculated. Thus words appearing in the title or specific fields [29] will have more weight than the ones presented in the body or raw contents.

A.2. Natural Language Processing. As Bohemia et al. [25] pointed out, keywords are usually formed by nouns, adjectives or present & past participles. This filtering technique consists on modifying the weight of a tag based on its morphosyntactic type.

Firstly, for each keyword, its Part Of Speech (POS) tag is calculated. Based on it, adjectives, nouns, and participles are stored according to their grammatical category. A small set of blacklisted POS tags are then removed while the rest continue to next step. Finally the value of each tag is increased based on the vector in which they are stored.

A.3. Clustering. The combination of *Normalization* and *NLP* techniques which has been previously described, provides a fast and simple method to suggest single-word tags. In some cases, unigrams cannot provide enough semantics, due to the simplicity of the meaning which can be supplied by an isolated word. As a consequence the use of n-grams, more specifically bigrams, or even short sentences are techniques that can be applied to improve the potential final set of suggestions.

The first step to generate n-grams consists in splitting the text into clusters from 1 to κ tokens. For each cluster, POS tags are generated. Next, they are analysed from the edges, removing non-suitable tokens, until a valid one is found. This process is performed once for each edge, removing the whole cluster if all of the tokens are no-suitable. Finally, as in the normalization for each cluster the number of occurrences is counted, multiplying this number for a weight.

B. Enrichment Stage. The enrichment stage receives a vector of tags as payback from the extraction stage. In this stage, tags are enriched by making queries to both ontologies and folksonomies.

B.1. Querying to Ontologies. This technique consists on enriching tags obtained by suggesting synonyms querying to the WordNet [30] dictionary, version 3.0.

Each tag is looked up in WordNet, if the tag is in the dictionary, its synonyms are retrieved, if not, the stage is finished. Each synonym is checked if it is within the vector of tags, if this happens, the tag's weight is increased, otherwise the synonym is added to the vector of tags with a default value.

B.2. Query to Folksonomies. As previous sections have presented, querying folksonomies can be a double-edge sword, on the one hand this technique increases the amount of representative words; on the other hand it can be unsafe because it can easily increase the amount of noise in the system. Taking into account this critical point, the presented methodology uses a chain of filters to only enrich that words over a given percentile (this parameter has as default value of 50 in order to skip the 50% less representative tags but it can be customized by the user in a later feedback stage). Therefore, only representative tags are enriched, skipping those that do not provide an extra meaning. This filtering process, indirectly, helps to reduce the number of queries to external services, usually REST calls, reducing bandwidth consumption and execution time.

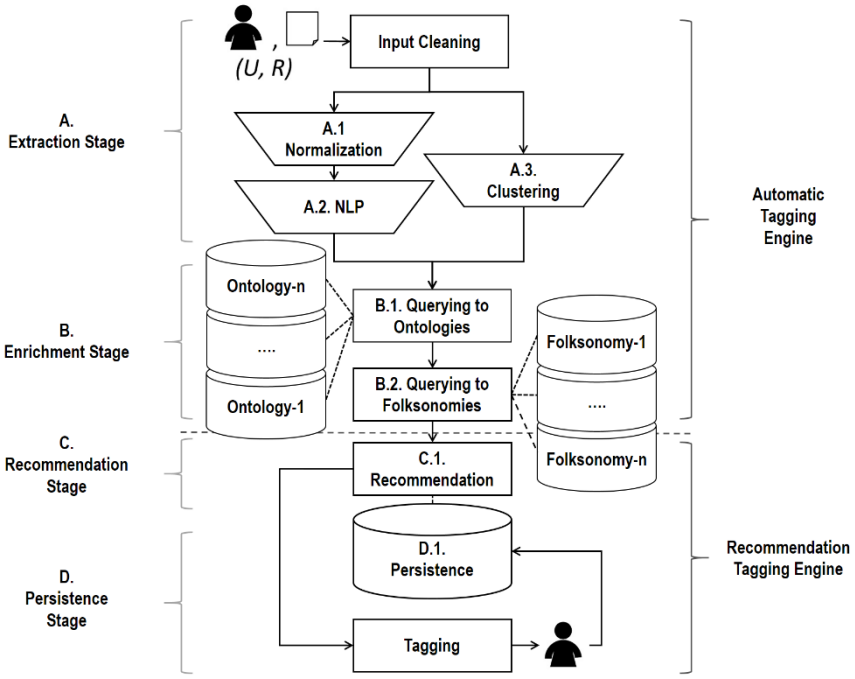


Fig. 1. Workflow of a Multilingual - Domainless Automatic Collaborative Tagging Methodology

3.2 Recommendation Tagging Engine

The main objective of the *Recommendation Tagging Engine*, is to suggest tags based on the previous behaviour of the users within the system.

C. Recommendation Stage. A well-known item-based recommender algorithm has been selected as technique for making suggestions. This algorithm recommends tags based on how similar items are to items [31], employing as similarity measure the Tanimoto Coefficient [32][33]. Furthermore the methodology has been designed to easily support the addition of new algorithms being technologically independent of the recommendation technique.

To make a short summary, the idea relies on: if a set of tags, for instance "wikipedia" and "encyclopaedia" are tagged together on several documents, the recommendation engine would suggest "encyclopaedia" if "wikipedia" is presented and vice versa.

The main drawback of this approach is that when the system is "cold" [17], the recommendation engine cannot work properly. In a pure recommendation engine, this can become a serious problem due to the sole reliance on these techniques. That is why a "*hybrid*" approach combining the automatic tagging engine with the recommendation tagging engine has been designed. Therefore, when the suggestion engine is cold, the automatic tagging engine still works underneath, providing valid and accurate results.

Example. The word "canary" has at least three different meanings, a domestic and colourful bird, the Spanish Archipelago in the northwest coast of Africa and a development version of the popular web browser, chrome[2]. In different scenarios, the recommendation engine should behave suitably to them, recommending tags based on their meaning according to the system's context. For instance in an e-learning system used for biology lessons, it would recommend tags related to that field, such as "bird", "animal" or "pet". In a travel agency, it would recommend tags like "islands", "tourist" or "Spain" and finally in a technology blog it would suggest tech tags such as "Chrome", "web browser" or "Google".

D. Persistence Stage. Once a set of tags is presented to the final user, she is able to decide whether to pick some of the provided tags or to add a new ones that are stored as feedback for future recommendations.

4 Overview of the architecture to support a methodology for multilingual and domain-less collaborative tagging. The ACOTA library.

This section outlines, a big picture, of the reference architecture of the proposed methodology. This reference implementation is called *ACOTA*, acronym of *Automatic Collaborative Tagging*. It is comprised by two main components, the *Core Component* which provides the *Automatic Tagging Engine*, and the *Feedback Component*, which provides the "collaborative" and "recommendation" capabilities. Both, *Core* and *Feedback* components can be used as a standalone projects or can be combined together to take advantage of the features provided by each module.

4.1 Core Component

This component enables and supports the *Automatic Tagging Engine*, it includes a set of *Enhancers* which provides features as terms extraction, natural language processing and querying to both ontologies and folksonomies.

A. Extraction Stage. The extraction stage is performed by the combination of *LuceneEnhancer*, *OpenNLPEnhancer* and *TokenizerEnhancer* implementing *Normalization*, *NLP* and *Clustering* phases, respectively.

A.1. Normalization. LuceneEnhancer implements the normalization stage and it extracts unigrams from an information resource, normalising the result. This *Enhancer* is implemented though the widely used Apache Lucene [34] in order to process the input text, removing stop-words and punctuation, among others.

[2] https://www.google.com/intl/en-419/chrome/browser/canary.html

A.2. Natural Processing Language. The *Natural Processing Language* phase is implemented by *OpenNLPEnhancer*. It uses the Apache OpenNLP [35] library to translate words into POS tags. This POS tags are used in order to apply filtering techniques which modifies the tag's weight, based on the morphosyntactic type of the tag.

A.3. Clustering. *TokenizerEnhancer* implements the *Clustering* phase, it employs advance natural language processing and extracting techniques in order to extract n-grams. As POS tag translator, this enhancer also implemented on the top of the Apache OpenNLP library.

B. Enrichment Stage. This stage consist in making queries to ontologies and folksonomies, in order to enrich the set of tags provided by the extraction stage. This stage has been using two enhancers based on Wordnet (*WordnetEnhancer)* and the Google Complete API (*GoogleEnhancer).*

B.1. Querying to Ontologies. It requires the WordNet dictionary to provide synonyms for English tags. It employs the Java Wordnet Interface (JWI) [36] capabilities to mount dictionary files directly on-disk, thus English tags are processed by these Enhancer in an isolated mode. Due to the flexibility of this implementation the support for new languages can be easily added configuring the Wordnet dictionary.

B.2. Querying to Folksonomies. It uses Google Complete[3] service as folksonomy, enriching the results with new tags. This folksonomy, suggest the top ten most used queries on Google, by a given set of words, in our case the tag to enrich is used as the query.

[3] https://www.google.es/

Table 1. Mapping between the proposed methodolgy and ACOTA

Phase	Stage	ID	ACOTA Component
Automatic Tagging engine	Extraction	A.1	Core
Automatic Tagging engine	Extraction	A.2	Core
Automatic Tagging engine	Extraction	A.3	Core
Automatic Tagging engine	Enriching	B.1	Core
Automatic Tagging engine	Enriching	B.2	Core
Recommendation Tagging Engine	Recommendation	C.1	Feedback
Recommendation Tagging Engine	Persistence	D.1	Feedback

4.2 Feedback Component

This component corresponds with the *Recommendation Tagging Engine* and it is divided in two sub-components, a recommendation engine which allows to recommend tags based on the previous users behaviour within the system, and the persistence system, which stores the users' feedback, enabling improvements in the accuracy of the recommendation engine.

C.1. Recommendation Engine. The LabelRecommenderEnhancer implements the Recommendation Stage, it suggest tags based on the previous users behaviour. This Enhancer is implemented on the top of Apache Mahout [37], a machine learning library which includes a predefined set of algorithms. The current prototype uses a customized version of the Mahout's item-based recommenders, with the aim of taking advantage of the previous effort and focussing on recommendations. It employs documents as users and tags as items.

D.1. Persistence. Existing machine learning libraries requires a persistence system based on different formats. In the case of Mahout previous data is stored in a relational database, more specifically in a table containing the tuple (document, item, weight). Furthermore the system has been designed to avoid database-lock in and other vendors or storage systems can be easily plugged to Mahout such as MySQL[38], MariaDB [39] and PostgreSQL[40] and also for the nosql dbms, MongoDB[41].

4.3 ACOTA in action

There is an available demo of ACOTA running at Heroku[4], see Fig 2, in the following URL **http://acota.herokuapp.com/**. This demo generates a set of tags from a body and a title inserted by the user. It also enables the modification of parameters

[4] https://www.heroku.com/

used by default. In addition to this, the web site includes information and tutorials about how to use ACOTA as a Java library or a REST service. It is also important to emphasize that it is an open source project under the Apache 2.0 license and can be located in the next GitHub[5] repositories: https://github.com/weso/acota-{component} (where component is: *core*, *seed*, *feedback* and *utils*).

5 Research Study

5.1 Design of the experiment

In order to evaluate the results we use two different evaluation criteria, precision and recall which was proposed by Cleverdon at 1966 [42]. Precision is the fraction of retrieved tags that are relevant, in this case, the fraction of proposed tags which matches with the tags tagged by the real users.

$$Precision = \frac{Match}{|\{Recommended\ Tags\}|} = \frac{|\{Dataset\ Tags\} \cap \{Recommended\ Tags\}|}{|\{Recommended\ Tags\}|}$$

Recall is the fraction of relevant tags that are successfully retrieved.

$$Recall = \frac{Match}{Count} = \frac{|\{Dataset\ Tags\} \cap \{Recommended\ Tags\}|}{|\{Relevant\ Tags\}|}$$

As request of our partner, in both evaluation criteria we have employed the top-12 tags from the suggestions vector (precision@12 and recall@12). Our partner, following the approach in [19], assumed that if users do not find an accurate tag within the top-12 tags, they would skip the search and probably they would end up self-adding a new one.

5.2 Sample

As an experimental dataset we have used a private dataset comprised by a set of 483 tagged documents and 1049 tags. This dataset has been provided in the context of a research project by the partner Treelogic S.L. It is a slice of its internal production corporate know–how management system and it has been generated by its workers in daily activities. These documents are in both English and Spanish, however there is a bigger number documents in Spanish than in English. These information resources are in a variety of lengths and formats (e.g. text plain, HTML & XML) including typical data heterogeneities (e.g. special characters). The data was not pre-processed or cleaned trying to simulate the whole real scenario.

[5] https://github.com

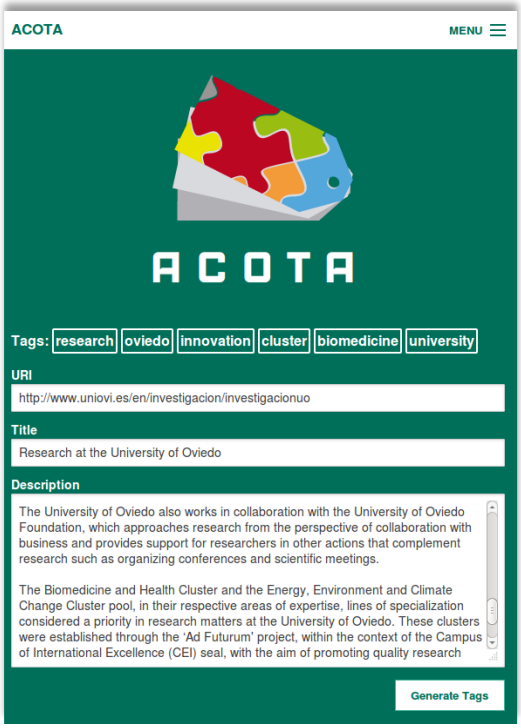


Fig. 2. A snapshot of the ACOTA web application demonstrator.

5.3 Results and Discussion

The precision and recall statistics are presented in **Table 2** and **Table 3**. The statistics are divided into three columns. Firstly, precision@12 is located in the second column, while recall@12 in the third one, respectively.

Extraction Stage As it was expected, the results show that the combination of Normalization + NLP and Clustering obtains the highest precision and recall values. Since the test data is comprised by unigrams and n-grams and these techniques are focused on recommending unigrams and n-grams respectively. The whole set of tags cannot be matched by each separated technique, due to the composition of the test data.

Table 2. Extraction Stage precision@12 and recall@12.

Heading level	Precision@12	Recall@12
1. Normalization	0.3737	0.0676
2. Normalization + NLP	0.3766	0.0682
3. Clustering	0.1144	0.0207
Extraction Stage	**0.4109**	**0.0744**

Enrichment Stage. The provided dataset contains a small percentage of English documents, as a consequence, querying to ontologies has almost no effect on the results. In contrast, querying to folksonomies increases dramatically both precision and recall.

Table 3. Enrichment Stage precision@12 and recall@12.

Heading level	Precision@12	Recall@12
4. Ontology Querying	0.4156	0.0752
5. Folksonomy Querying	0.4290	0.0776
Enrichment Stage	**0.4290**	**0.0776**

It is worth mentioning that the test data contain a certain proportion of tags which make reference to temporal events, internal projects or even personal matters. Therefore, the number of measurable tags are slightly reduced, and even a folksonomy beyond measure as Google Complete API is not able to suggest these unmatched tags.

As it can be appreciated in both tables, recall@12 has low values. Due to the fact that the dataset-provided contains 1049 tags for 483 documents, the average of tags per documents is near 2.1718. Thus, the ideal recall@12 value would be 0.1810 (1049 tags tagged by experts per 5796 tags suggested by our system, 12 by each document). Therefore, the recall values are quite accurate, reaching values of 42.90%.

6 Conclusions and Future Work

We have proposed a methodology for tagging multilingual documents in a semiautomatic way, employing extraction, enrichment and user-behaviour recommendation techniques. Our methodology transfers the final decision of choosing the proper tag to the user, so if the recommended tags are non-suitable to him, he would add a new one feedbacking the system. This feedback allows the recommendation engine to improve the results based on the previous behaviour of the users.

We have also developed a reference implementation, ACOTA, which was tested against production data from a technology company. Despite the considerable amount of internal and temporal tags, valid results have been obtained. This methodology and its implementation help knowledge workers to minimize the categorization-act effort providing a tool a better information resources classification within a knowledge organization.

As future work, we have considered increasing the amount of native languages, since English and Spanish are now supported, new languages must be added in order to internationalize the methodology and ACOTA as much as possible. Furthermore the experimentation with new large and diverse datasets will give the proper feedback to improve the accuracy of each component. Another issue that we have taken into account is to improve the performance of the system. Although the system has a good performance with regular documents, it suffers when real-time (in terms of seconds) suggestions must be done processing large documents that contain several thousands of words.

Acknowledgements

ACOTA is a subproject of the ORIGIN project, co-financed by ERDF (European Regional Development Fund), with the aim of increasing the productivity of software development activities in global scenarios.

References

1. Belkin, N., Croft, W.: Information filtering and information retrieval: two sides of the same coin? Communications of the ACM. 29, 1–10 (1992).
2. What is big data?, http://www-01.ibm.com/software/data/bigdata/.
3. Shklar, L., Sheth, A., Kashyap, V., Shah, K.: InfoHarness: Use of automatically generated metadata for search and retrieval of heterogeneous information. Advanced Information Systems Engineering. pp. 217–230 (1995).
4. Large, A., Moukdad, H.: Multilingual access to web resources: an overview. Program: electronic library and information systems. 34, 43–58 (2000).
5. W3Techs: Usage of content languages for websites, http://w3techs.com/technologies/overview/content_language/all.
6. Kunder, M. de: The size of the World Wide Web (The Internet), http://www.worldwidewebsize.com/.
7. Hjørland, B.: Semantics and knowledge organization. Annual Review of Information Science and Technology. 41, 367–405 (2007).
8. Hjørlandd, B.: Methods for evaluating information sources: An annotated catalogue. Journal of Information Science. 38, 258–268 (2012).
9. Gruber, T.: Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies. 43, 907–928 (1995).
10. Wal, T. Vander: Folksonomy Coinage and Definition, http://vanderwal.net/folksonomy.html.
11. Casado-Lumbreras, C., Rodríguez-González, A., Álvarez-Rodríguez, J.M., Colomo-Palacios, R.: PsyDis: Towards a diagnosis support system for psychological disorders. Expert Systems with Applications. 39, 11391–11403 (2012).
12. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. The Semantic Web: Research and Applications. 4011, 411–426 (2006).
13. Yoo, D., Choi, K., Suh, Y., Kim, G.: Building and evaluating a collaboratively built structured folksonomy. Journal of Information Science. (2013).
14. Shirky, C.: Ontology is Overrated: Categories, Links, and Tags, http://www.shirky.com/writings/ontology_overrated.html?goback=.gde_1838701_member_179729766.
15. Halpin, H., Robu, V., Shepard, H.: The dynamics and semantics of collaborative tagging. Proceedings of the 1st Semantic Authoring and Annotation Workshop SAAW'06. (2006).
16. Newman, M.: Power laws, Pareto distributions and Zipf's law. Contemporary physics. (2005).
17. Park, S.-T., Pennock, D., Madani, O., Good, N., DeCoste, D.: Naïve filterbots for robust cold-start recommendations. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06. pp. 699–705. ACM Press, New York, New York, USA (2006).
18. Brook, C.H., Montanez, N.: Improved annotation of the blogopshere via autotagging and hierarchical clustering. Proceedings of the 15th World Wide Web Conference (WWW06). (2006).
19. Mishne, G.: AutoTag. Proceedings of the 15th international conference on World Wide Web (WWW 06). p. 953. ACM Press, New York, New York, USA (2006).

20. Sood, S.C., Owsley, S.H., Hammond, K.J., Birnbaum, L.: TagAssist: Automatic Tag Suggestion for Blog Posts. (2007).
21. Noll, M.G., Au Yeung, C., Gibbins, N., Meinel, C., Shadbolt, N.: Telling experts from spammers. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09. p. 612. ACM Press, New York, New York, USA (2009).
22. Hong, L., Ahmed, A., Gurumurthy, S.: Discovering geographical topics in the twitter stream. Proceedings of the 21st international conference on World Wide Web. 769–778 (2012).
23. Gayo-avello, D.: blindLight: Una nueva técnica para procesamiento de texto no estructurado mediante vectores de n-gramas de longitud variable con aplicación a diversas tareas de tratamiento de lenguaje natural, (2005).
24. Mika, P., Ciaramita, M., Zaragoza, H., Atserias, J.: Learning to Tag and Tagging to Learn: A Case Study on Wikipedia. IEEE Intelligent Systems. 23, 26–33 (2008).
25. Dostal, M., Ježek, K.: Automatic keyphrase Extraction extraction based on NLP Automatic Keyphrase based on NLP and statistical methods and Statistical Methods. Proceedings of the Dateso 2011: Annual International Workshop on DAtabases, TExts, Specifications and Object. 140–145 (2011).
26. Summary by language size, http://www.ethnologue.com/statistics/size.
27. Jimenez-Nácero, W., Luis-Alvargonzález, C., Abella-Vallina, P., Alvarez-Rodríguez, J.M., Labra-Gayo, J.E., Ordoñez de Pablos, P.: Emergent Ontologies by collaborative tagging for Knowledge Management. Advancing Information Management through Semantic Web Concepts and Ontologies. p. 16. IGI-Global (2012).
28. Kern, R., Granitzer, M., Pammer, V.: Extending Folksonomies for Image Tagging. 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services. pp. 126–129. IEEE (2008).
29. Suchanek, F.M., Vojnovic, M., Gunawardena, D.: Social tags. Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08. p. 223. ACM Press, New York, New York, USA (2008).
30. Miller, G. a.: WordNet: a lexical database for English. Communications of the ACM. 38, 39–41 (1995).
31. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. Proceedings of the tenth international conference on World Wide Web - WWW '01. pp. 285–295. ACM Press, New York, New York, USA (2001).
32. Lipkus, A.: A proof of the triangle inequality for the Tanimoto distance. Journal of Mathematical Chemistry. 26, 263–265 (1999).
33. Tanimoto, T.T.: An elementary mathematical theory of classification and prediction. (1958).
34. A. Jakarta: Apache Lucene-a high-performance, full-featured text search engine library, http://lucene.apache.org/core/.
35. Baldridge, J.: The opennlp project, http://opennlp.apache.org/, (2005).
36. Finlayson, M.A.: JWI: The MIT Java Wordnet Interface, http://projects.csail.mit.edu/jwi/, (2008).
37. Foundation, A.: Apache Mahout: Scalable machine learning and data mining, http://mahout.apache.org, (2008).
38. DuBois, P.: MySQL administrator's guide. (2004).
39. Foundation, M.: MariaDB: An enhanced, drop-in replacement for MySQL., https://mariadb.org/, (2009).
40. Momjian, B.: PostgreSQL: introduction and concepts. (2001).
41. K Chodorow, M.D.: MongoDB: the definitive guide. O'Reilly Media (2010).

42. Cleverdon, C., Mills, J., Keen, M.: Factors determining the performance of indexing systems. (1966).

Assisting Corporate Knowledge Management: A Multilingual and Domainless Semi-Automatic Collaborative Tagging Approach

César Luis Alvargonzález[a], Jose María Álvarez-Rodríguez[b] and Jose Emilio Labra Gayo[a]

[a] WESO Research Group, Department of Computer Science, University of Oviedo,
33005, Oviedo, Spain
cesar.luis@weso.es, labra@uniovi.es
[b] South East European Research Center, 54622, Thessaloniki, Greece
jmalvarez@seerc.org

Abstract. Despite of automatic, social and collaborative tagging techniques have been the topic of certain number of recent studies, in the meantime, multilingual tagging techniques has been poorly noticed.

The present paper introduces a multilingual hybrid methodology to automatically deploy and combine collaborative tagging techniques based on user-behaviour and well-known recommendation algorithms to name a few. A reference implementation called ACOTA (Automatic Collaborative Tagging) is also outlined in order to demonstrate the novel recommendation capabilities of this approach with the aim of assisting both expert and non-expert users when multilingual resource tagging is required. Finally a quantitative research study in the context of corporate knowledge management is also presented to evaluate and assess the goodness and accuracy of the methodology to minimize the effort of multilingual document categorization.

Keywords: Collaborative Tagging; Automatic Tagging; Multilingual data; Document Retrieval; Knowledge Management

1 Introduction

The sheer mass of data already available in the Internet and the increasing use of mobile devices such as smart-phones, tablets or e-books readers are generating a new and dynamic data/information domain in which new requirements are emerging [1]. Some time ago information resources were printed on paper and kept in cabinets but nowadays this new vast amount of data is commonly stored in digital formats using different organization and access techniques with the objective to facilitate processes such as information and document retrieval or search and report, to name a few. Although they try to take advantage of indexing and categorizing algorithms, on-line information resources usually contain a lot of heterogeneities [2] that cannot be easily addressed. More specifically, the multilingual encoding of information is one of the main drawbacks to provide a common access to the information [3]. In this sense, Internet contents are currently available on a huge variety of languages but English is the most common language [4], used in 54.9% of websites, and most of the tools and techniques to perform some kind of exploitation over this information are customized to work on it. This implies that the rest of information (45.1%) faces a lack of tools to properly exploit this information. Taking into account that there are around 14.24 billion indexed web sites [5], 6.44 billion of web sites are in other languages rather than English, and as a consequence new tools are required to bridge the gap between the exploitation capabilities of English and non-English information resources.

Furthermore organizing this vast amount of multilingual data can be tedious, but in some enterprise and academics fields, it can become the cornerstone to success. E-learning systems, B2B environments, Corporate Knowledge Management or Extraction and Information Retrieval are some domains in which data filtering and management processes are becoming crucial to get more accurate and timely information with the aim of enabling new business opportunities. More specifically knowledge organizations [6] and workers have gained their momentum in the new information society. Activities such as efficient document description, indexing or classification of information resources are now major challenges due to the aforementioned dynamic data domain. In this sense Hjørland [7] establishes different approaches such as the use of controlled domain-based vocabularies and information retrieval techniques, user-oriented cognitive views or bibliometric approaches among others in order to ease the tasks of knowledge workers. As a consequence organizations with the capability of managing their knowledge activities can take advantage of discovering new business opportunities or workers capabilities due to the analysis of their daily activities. In this sense technology watching is one of the key activities in some of these companies (as an example Treelogic S.L.[1], a technology company, has established a percentage of the working time for this task) in which workers add everyday news, blog posts, research articles, funding opportunities, etc. that help companies to be aware of new trends. In this environment the proper

[1] http://www.treelogic.com

classification and exploitation of information resources leads to a knowledge organization that can exploit this information and, as a consequence, its know-how for improving their own performance.

In order to address these new requirements in knowledge organizations there are several well-known and traditional techniques to model, structure, organise and exploit data and information such as conceptual maps, taxonomies or controlled vocabularies. Nevertheless recent times have shown a growing use of ontologies and folksonomies as a method to efficiently manage big amount of information resources. An ontology is usually defined as a specification of a shared conceptualization [8], it is a formal description of concepts and their relationships involved in knowledge domain that assist to organize and build knowledge-based system. On the other hand, a folksonomy [9] is the result of free categorization, without a defined structure or formality. Usually it is created by collective intelligence; a group of users with different levels of knowledge interested in some domain collect and classify information resources. These two emerging approaches have been successfully applied to specific domains such as e-Health, e-Government or e-Procurement (ontologies [10][11][12]) and Web 2.0 sites (folksonomies [13][14][15][16][17][18]) such as Delicious[2], Flickr[3], Youtube[4], Linkedin, Twitter[5] or Facebook[6].

In general, ontologies are more adequate to classify domain-specific information and data in a restricted context in which experts in a field have reached a common and shared understanding. In this context, data, information and knowledge is commonly concrete, static and stable. Obviously this approach presents some drawback to tackle data, information and knowledge management in a more general and dynamic environment. Nevertheless some structural upper ontologies such as DOLCE[7], SUMO[8] or PROTON[9] have been delivered in order to formalize high-level entities and relationships. On the other hand, as Shirky pointed in [19] a folksonomy works better with a large and dynamic corpus, unstable entities and participants of all levels of knowledge that seem to fit better in the web trying to exploit information to provide business intelligent services such as marketing suggestions or knowledge discovery. Furthermore another issue to take into account in data and information management lies in user behaviour, some works [20], have demonstrated that once a tag has been included within the system, there is a high probability that it will be used again, obeying the Pareto principle rule 80-20 [21].

According to the aforementioned key points: a growing and dynamic new data domain needs, leads to knowledge organizations to use ontologies and folksonomies as knowledge management approaches. A hybrid methodology is presented employing automatic tagging techniques and user-behaviour recommendation algorithms to take advantage of folksonomies as a previous step to consolidate knowledge in an ontology. This methodology is built on top of different techniques as natural language processing, querying to both folksonomies and ontologies, collaborative or machine learning techniques among others to deliver a multilingual methodology which does not suffer from cold start [22]. We have developed a software library called Automatic COllaborative Tagging (hereafter ACOTA) as a reference implementation that provides an automatic tagging engine with collaborative and user-behaviour recommendation capabilities. A quantitative evaluation taking into account existing information retrieval measures such as precision and recall is also performed over a real dataset (480 tagged information resources) created by the employees of the Treelogic S.L. company with the objective of assessing the tagging capabilities of the ACOTA library.

The rest of the paper is structured as follows. Section 2 describes related work. Section 3 presents the proposed methodology. An overview of the architecture to support a methodology for multilingual and domain-less collaborative tagging is in the section 4. Section 5 presents a knowledge-management case study. Section 6 presents the experiment methodology, evaluation and the discussion and finally section 7 we present some conclusions.

[2] https://delicious.com/
[3] http://www.flickr.com/
[4] http://www.youtube.com/
[5] https://twitter.com/
[6] https://www.facebook.com/
[7] http://www.loa.istc.cnr.it/DOLCE.html
[8] http://www.ontologyportal.org/
[9] http://proton.semanticweb.org/

2 Related Work

Taking into account the use of blogs as information sources there are a number of projects [23][24][25] trying to generate tags from these sources. Although information knowledge goes beyond blog-post tagging, it has a certain similarity, since a blog-post can be seen as traditional document, that cover many and specific different topics, rather than web sites. In this sense, Brooks et al. [23] built a system that was based on the use of the top three term frequency–inverse document frequency (TFIDF) score tags of a blog post. This approach groups documents into clusters with the hypothesis that a cluster of documents that shared a tag should be more similar than a randomly constructed set of documents. The proposed methodology uses a similar approach but with a different purpose, to suggest for a given tag, common tags that are within the same cluster. AutoTag developed by Mishne et al. [24] is a tag suggestion blog-post engine which employs collaborative filtering methods. This tool offers suggestions for tags based on tags assigned to similar posts, leaving the editor the decision of choosing the proper set of tags. Our methodology also transfers the final decision to the user, and employs similar suggestion techniques, but it does not rely only on collaborative methods, adding a pre-automatic tag generation step, based on the structure of the document. Finally, TagAssist is another automatic tag suggestion engine developed by Sood et al. [25], it evolves AutoTag design, adding support for tag compression. This approach provides a good support for English posts that serves as inspiration for the proposed multilingual methodology but it is exclusive for this language, in contrast with our solution. Furthermore and due to the growing use of social networks, some works have emerged to detect experts [26], trends and filter information [27] among others. The approach here presented takes advantage of these existing work to adapt broad techniques in a narrower context such as knowledge management in a company.

In the case of Automatic Keywords Extraction [28][29], Song et al. [30] proposed a clustering and classification based tag recommendation system. Sun et al. [31] evolved it with a language model for tag recommendation (LMTR) approach, this technique compares how similar documents are to documents based on the shared words. Our research employs a slightly different approach than these works, it also compares how similar are documents but instead, taking into account the shared tags (previously selected by the users) by the documents. Dostal et al. [32] developed an automatic keyphrase extraction based on Natural Language Processing (NLP) and statistical methods. In this paper we employ more customized and domain-based NLP techniques.

Most of the existing works are focused on English tag recommendation. In the case of TagAssist [25] and Brooks [23] they drop out any non-English words; on the other hand AutoTag [24] takes a less aggressive solution, holding non-English words, but due to the application of an English stemmer for any word, regardless of the language, it ends up giving low scores to non-English posts. In the case of ACOTA, as a multilingual automatic and collaborative tagging engine, it is focused on providing support for non-English languages, e.g. Spanish due to the motivating scenario, apart from English. That is why one of our initial design requirements was to facilitate the configuration and the extension to other languages.

Another issue in recommendation systems is the cold start [22]. It appears when the system is relatively new or unused, so there is not enough data to properly perform the recommendation algorithms. This issue is partially addressed by the methodology and ACOTA and it is able to suggest comprehensive tags even when there is no data stored on the system given that part of the suggestion takes as input public existing folksonomies. In this sense, a similar approach has been followed in previous works [33][34][35][36] that have been made use of queries to folksonomies such as Google Complete API to enrich tag suggestion. Nevertheless in the context of the present paper results of empirical experiments show a considerable amount of noise that decreases the accuracy of suggestions. For instance the delegation of suggestions to this service without any pre/post processing implies the generation of a vector of suggestions with n tags, where some tags are representative according to the initial query but they are not usually representative enough for a domain-based recommendation. In order to partially reuse this service and avoid non-representative tags our system pre-filter the queries and filters the external suggestions according to the existing domain.

Taking into account the main features and highlights of the aforementioned works, we have extracted the top features, which a domain less automatic collaborative tagging technique should have, see **Table 1**.

Table 1. Recommender Criteria

Feature
Multilingual
Collaborative
Domainless (No Specific Domain)
Not Suffering from Cold Start (No Training Required)
NLP Techniques
Tag Compression
Querying to Ontologies
Querying to Folksonomies

3 A Multilingual - Domainless Automatic Collaborative Tagging Methodology

In this section a detailed description of the proposed methodology is provided. The methodology and, as a consequence ACOTA, is divided into two main stages: 1) the automatic tagging engine that is in charge of text normalization, keyword extraction, and tag enrichment… 2) the recommendation tagging engine that uses the outcomes of the first stage to finally generate tag suggestions in a certain context.

3.1 Automatic Tagging Engine

The automatic tagging engine is comprised of two stages. Firstly, the extraction stage which consists on tag extraction based on the document structure is applied to an existing dataset. Afterwards, the previous results are enhanced in the enrichment stage, querying to both folksonomies and ontologies. As a result, a set of descriptive tags are generated by the automatic tagging engine. Before this process and as usual in any technique dealing with natural language, an optional but relevant step must be executed in order to clean the raw text, removing special characters, stop-words, etc. The aim of this cleaning is to decrease the potential noise and to avoid the spreading of these non-useful lemmas to further stages.

A. Extraction Stage. This stage extracts tags from a document. It is composed of two different extracting techniques: 1) unigrams generation, it is performed by normalization and NLP, and 2) n-grams generation, a refined combination of the aforementioned techniques.

A.1. Normalization. This phase takes as input an information resource and produces a vector of unigrams. The first step is to split words at punctuation characters, removing punctuation marks and other delimiters. Then, each token is turned to a lower case. Tokens with less than three characters are removed as they are usually irrelevant for the results [36][37]. Finally, tokens are filtered according to a stop-word set (depending on the language common stop-words sets are available and can be configured), ending up with a clean set of unigram tags.

For each tag the number of occurrences is counted, multiplying this number by a weight. According to the word position or appearance, the weight is calculated. Thus words appearing in the title or specific fields [38] will have more weight than the ones presented in the body or raw contents.

A.2. Natural Language Processing. As Bohemia et al. [32] pointed out, keywords are usually formed by nouns, adjectives or present & past participles. This filtering technique consists on modifying the weight of a tag based on its morphosyntactic type.

Firstly, for each keyword, its Part Of Speech (POS) tag is calculated. Based on it, adjectives, nouns, and participles are stored according to their grammatical category. A small set of blacklisted POS tags are then removed while the rest continue to the next step. Finally the value of each tag is increased based on the vector in which they are stored.

A.3. Clustering. The combination of *Normalization* and *NLP* techniques provides a fast and simple method to suggest single-word tags. In some cases, unigrams cannot provide enough semantics, due to the simplicity of the meaning which can be supplied by an isolated word. As a consequence the use of n-grams, more specifically bigrams, or even short sentences are techniques that can be applied to improve the potential final set of suggestions.

The first step to generate n-grams consists in splitting the text into clusters from 1 to κ tokens. For each cluster, POS tags are generated. Next, they are analysed from the edges, removing non-suitable tokens, until a valid one is found. This process is performed once for each edge, removing the whole cluster if all of the tokens are no-suitable. Finally, as in the normalization for each cluster the number of occurrences is counted, multiplying this number for a weight.

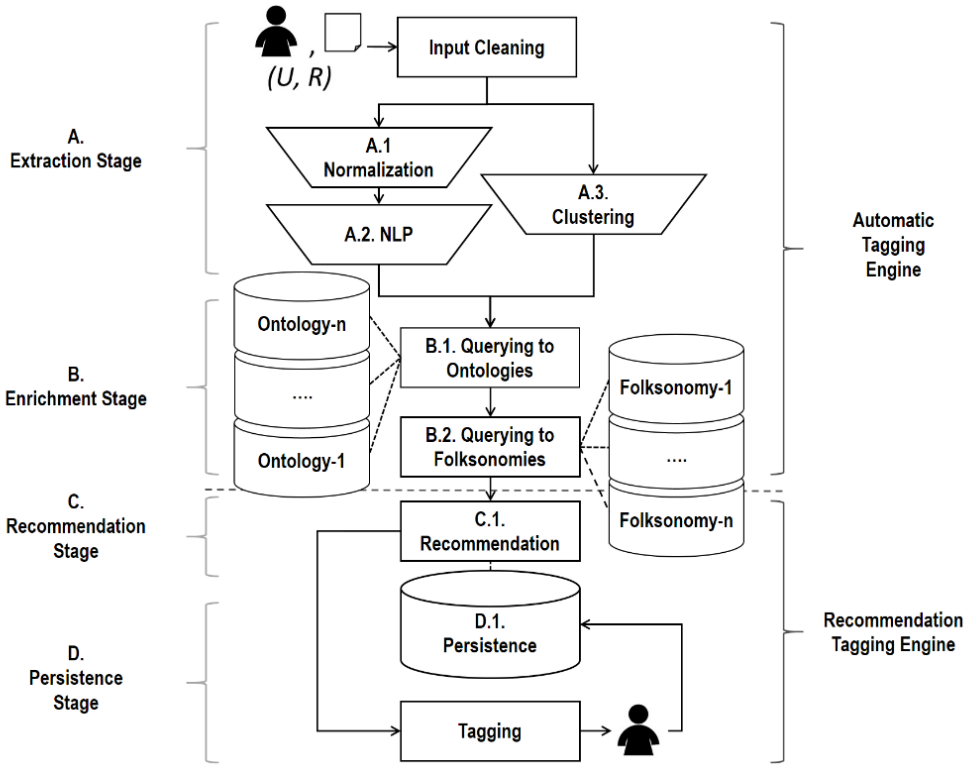


Fig. 1. Workflow of a Multilingual - Domainless Automatic Collaborative Tagging Methodology

B. Enrichment Stage. The enrichment stage receives a vector of tags as payback from the extraction stage. In this stage, tags are enriched by making queries to both ontologies and folksonomies.

B.1. Querying to Ontologies. This technique consists on enriching tags obtained by suggesting synonyms querying to the WordNet [39] dictionary, version 3.0.

Each tag is looked up in WordNet, if the tag is in the dictionary, its synonyms are retrieved, if not, the stage is finished. Each synonym is checked if it is within the vector of tags, if this happens, the tag's weight is increased, otherwise the synonym is added to the vector of tags with a default value.

B.2. Query to Folksonomies. As previous sections have presented, querying folksonomies can be a double-edge sword, on the one hand this technique increases the amount of representative words; on the other hand it can be unsafe because it can easily increase the amount of noise in the system. Taking into account this critical point, the presented methodology uses a chain of filters to enrich only those words over a given percentile (this parameter has as default value of 50 in order to skip the 50% less representative tags but it can be customized by the user in a later feedback stage). Therefore, only representative tags are enriched, skipping those that do not provide an extra meaning. This filtering process,

indirectly, helps to reduce the number of queries to external services, usually REST calls, reducing bandwidth consumption and execution time.

Improving performance: Caching REST Calls. An underlying problem in the on-going approach lies in the time consumption when external services are requested to enrich or extend some of the words. Due to the fact that a single tag enrichment requires, at least, one external REST call, when the number of tags is big enough, the execution time can be overkill, even with the filters employed by our approach.

In order to prevent this situation, a configurable cache has been added to anticipate and save previous results of REST requests, thus the system is able to avoid a big number of requests, reducing a lot of bandwidth consumption and execution time. Nevertheless there is a critical issue that it is under study and lies in the penalty time when the REST call is not cached, this happens as a result of the need to check the availability, perform the REST call and finally save it back to the cache.

Fig. 2. Ongoing Example (Top-12 Tags)

In the particular case of Intranets or enterprise environments, where the amount of processed documents is considerable, the use of a cache speed up dramatically the process, given that a huge amount of words will be stored within the cache, reducing the querying time to the same time of calling a local method. In contrast, specific or new anticipated enrichments, due to the lack of tags within the cache, can considerably reduce the performance of the enrichment process.

3.2 Recommendation Tagging Engine

The main objective of the *Recommendation Tagging Engine*, is to suggest tags based on the previous behaviour of the users within the system.

C. Recommendation Stage. A well-known item-based recommender algorithm has been selected as technique for making suggestions. This algorithm recommends tags based on how similar items are to items [40], employing as similarity measure the Tanimoto Coefficient [41][42]. Furthermore the methodology has been designed to easily support the addition of new algorithms being technologically independent of the recommendation technique.

To make a short summary, the idea relies on: if a set of tags, for instance "wikipedia" and "encyclopaedia" are tagged together on several documents, the recommendation engine would suggest "encyclopaedia" if "wikipedia" is presented and vice versa.

The main drawback of this approach is that when the system is "cold" [22], the recommendation engine cannot work properly. In a pure recommendation engine, this can become a serious problem due to the sole reliance on these techniques. That is why a "*hybrid*" approach combining the automatic tagging engine with the recommendation tagging engine has been designed. Therefore, when the suggestion engine is cold, the automatic tagging engine still works underneath, providing valid and accurate results.

Example. The word "canary" has at least three different meanings, a domestic and colourful bird, the Spanish Archipelago in the northwest coast of Africa and a development version of the popular web browser, chrome[10]. In different scenarios, the recommendation engine should behave suitably to them, recommending tags based on their meaning according to the system's context. For instance in an e-learning system used for biology lessons, it would recommend tags related to that field, such as "bird", "animal" or "pet". In a travel agency, it would recommend tags like "islands", "tourist" or "Spain" and finally in a technology blog it would suggest tech tags such as "Chrome", "web browser" or "Google".

D. Persistence Stage. Once a set of tags is presented to the final user, she is able to decide whether to pick some of the provided tags or to add a new ones that are stored as feedback for future recommendations.

4 Description of the ACOTA System

We developed a library called *ACOTA* (*Automatic Collaborative Tagging)*. It consists of two main components, the *Core Component* which provides the *Automatic Tagging Engine*, and the *Feedback Component*, which provides the "collaborative" and "recommendation" capabilities. Both, *Core* and *Feedback* components can be used as standalone projects or can be combined together to take advantage of the features provided by each module.

4.1 Core Component

This component enables and supports the *Automatic Tagging Engine*, it includes a set of *Enhancers* which provides features as terms extraction, natural language processing and querying to both ontologies and folksonomies.

A. Extraction Stage. The extraction stage is performed by the combination of *LuceneEnhancer*, *OpenNLPEnhancer* and *TokenizerEnhancer* implementing *Normalization*, *NLP* and *Clustering* phases, respectively.

A.1. Normalization. LuceneEnhancer implements the normalization stage and it extracts unigrams from an information resource, normalising the result. This *Enhancer* is implemented though the widely used Apache Lucene[11] in order to process the input text, removing stop-words and punctuation, among others.

[10] https://www.google.com/intl/en-419/chrome/browser/canary.html
[11] http://lucene.apache.org/

A.2. Natural Language Processing. The *Natural Language Processing* phase is implemented by *OpenNLPEnhancer*. It uses the Apache OpenNLP[12] library to translate words into POS tags. This POS tags are used in order to apply filtering techniques which modifies the tag's weight, based on the morphosyntactic type of the tag.

A.3. Clustering. TokenizerEnhancer implements the *Clustering* phase, it employs more advanced natural language processing and extraction techniques in order to extract n-grams. As POS tag translator, this enhancer has also been implemented using the Apache OpenNLP library.

B. Enrichment Stage. This stage consist in making queries to ontologies and folksonomies, in order to enrich the set of tags provided by the extraction stage. This stage has been using two enhancers based on Wordnet (*WordnetEnhancer)* and the Google Complete API (*GoogleEnhancer).*

B.1. Querying to Ontologies. It requires the WordNet dictionary to provide synonyms for English tags. It employs the Java Wordnet Interface[13] (JWI) capabilities to mount dictionary files directly on-disk, thus English tags are processed by these Enhancer in an isolated mode. Due to the flexibility of this implementation the support for new languages can be easily added configuring the Wordnet dictionary.

B.2. Querying to Folksonomies. It uses Google Complete[14] service as folksonomy, enriching the results with new tags. This folksonomy, suggest the top ten most used queries on Google, by a given set of words, in our case the tag to enrich is used as the query.

As it has been previously explained in the methodology, the prototype employs an optional caching system, in order to reduce bandwidth consumption and execution time. As caching system we have employed Memcached[15], a high-performance and distributed memory object caching system

Table 2. Mapping between the proposed methodology and ACOTA

Phase	Stage	ID	ACOTA Component
Automatic Tagging engine	Extraction	A.1	Core
Automatic Tagging engine	Extraction	A.2	Core
Automatic Tagging engine	Extraction	A.3	Core
Automatic Tagging engine	Enriching	B.1	Core
Automatic Tagging engine	Enriching	B.2	Core
Recommendation Tagging Engine	Recommendation	C.1	Feedback
Recommendation Tagging Engine	Persistence	D.1	Feedback

4.2 Feedback Component

This component corresponds with the *Recommendation Tagging Engine* and it is divided in two sub-components, a recommendation engine which allows the system to recommend tags based on the previous users behaviour within the system, and the persistence system, which stores the users' feedback, enabling improvements in the accuracy of the recommendation engine.

C.1. Recommendation Engine. The *LabelRecommenderEnhancer* implements the Recommendation Stage, it suggest tags based on the previous users behaviour. This Enhancer is implemented using Apache Mahout[16], a machine learning library which includes a predefined set of algorithms. The current prototype uses a customized version of the Mahout's item-

[12] http://opennlp.apache.org/
[13] http://projects.csail.mit.edu/jwi/
[14] https://www.google.es/
[15] http://memcached.org/
[16] http://mahout.apache.org/

based recommenders, with the aim of taking advantage of the previous effort and focussing on recommendations. It employs documents as users and tags as items.

D.1. Persistence. Existing machine learning libraries requires a persistence system based on different formats. In the case of Mahout previous data is stored in a relational database, more specifically in a table containing the tuple (document, item, weight). Furthermore the system has been designed to avoid database-lock in and other vendors or storage systems can be easily plugged to ACOTA such as MySQL[17], MariaDB[18] and PostgreSQL[19] and also for the NOSQL systems like MongoDB[20].

4.3 ACOTA in action

There is an available demo of ACOTA running at Heroku[21], see **Fig. 3**, in the following URL **http://acota.herokuapp.com/**. This demo generates a set of tags from a body and a title inserted by the user. It also enables the modification of parameters used by default. In addition to this, the web site includes information and tutorials about how to use ACOTA as a library or a REST service. It is also important to emphasize that it is an open source project under the Apache 2.0 license and can be located in: https://github.com/weso/acota-{component} (where component is: *core*, *seed*, *feedback* and *utils*).

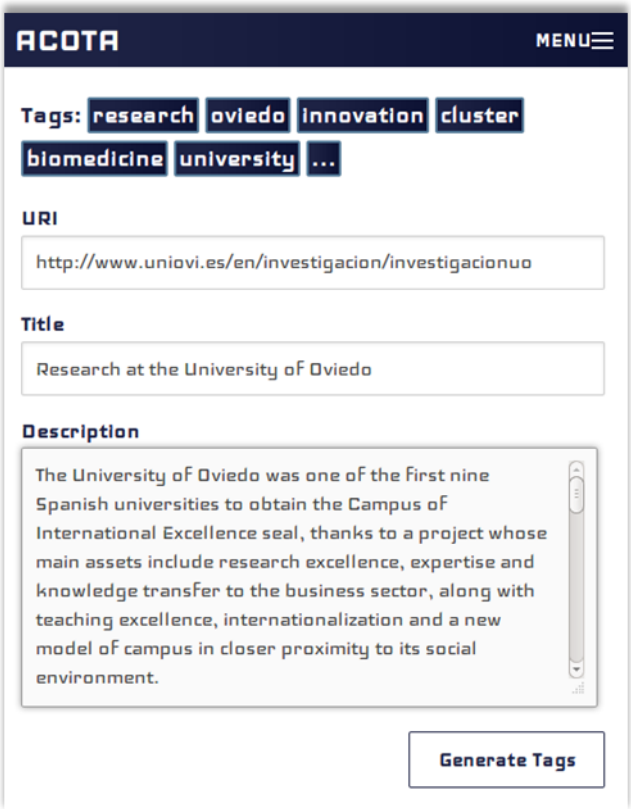


Fig. 3. Screen capture of ACOTA demo.

[17] http://www.mysql.com/
[18] http://mariadb.org/
[19] http://www.postgresql.org/
[20] http://www.mongodb.org/
[21] https://www.heroku.com/

5 A Corporate Knowledge Management case of study

Treelogic S.L. is a Spanish company which provides customers with information and communication technology-based solutions. Treelogic S.L. has a solution called *Imaginn Watching*[22], a research and development (R&D) technological watching tool. The aim of this solution is to be up to date with the latest trends in technology and business opportunities within their sector.

Technological watch consists [43] in "watching" regularly similar areas such as legal, social, technological or environmental in order to have the company's internal information up to date and ready to be consulted by the company's decision-makers.

This kind of tool requires handling and organising vast amounts of data. This data usually comes from the Internet sources in a variety of languages. The domain of the data is weak and is constantly changing as new terms, technologies and business opportunities are coined regularly [44]. All this information has to be created and consumed by workers from different company's areas in a collaborative way.

These technological watch requirements fit with the multilingual-domain less automatic collaborative tagging features aforementioned in section 2. Multilingual support, collaborative capabilities, the possibility to enrich the data with external sources of information (such as folksonomies, ontologies and so forth) and the lack of domain specific, cold start and the necessity of training the tagging engine, are the requirements that Treelogic S.L. have looked to fulfil with ACOTA.

According to **Table 3** our proposal is compared to other approaches, we do not evaluate the details of the implementation of each solution, and we only focus on some general properties. Our approach fits all the features with the exception of its main drawback, the lack of tag compression, which would help to improve the results. On the other hand, since tag compression is tightly bounded to a specific language, it was dismissed.

Table 3. Collaborative Tagging Technique for assisting corporate knowledge management criteria

Feature	ACOTA	Brooks	AutoTag	TagAssist	Bohemia
Multilingual	+	-	-, partial	-	-
Collaborative	+	+	+	+	+
Domainless (No Specific Domain)	+	-, Requires Training	-, Requires Training	-, Requires Training	+
Not Suffers from Cold Start (No Training Required)	+	-	-	-	-
NLP Techniques	+	-	-	+	+
Tag Compression	-	-	-	+	-
Querying to Ontologies	+	-	-	-	-
Querying to Folksonomies	+	-	-	-	-

6 Research Evaluation

6.1 Design of the experiment

In order to evaluate the results we use two different evaluation criteria, precision and recall which was proposed by Cleverdon [45]. Precision (1) is the fraction of retrieved tags that are relevant, in this case, the fraction of proposed tags which matches with the tags tagged by the real users.

$$Precision = \frac{Match}{|\{Recommended\ Tags\}|} = \frac{|\{Dataset\ Tags\} \cap \{Recommended\ Tags\}|}{|\{Recommended\ Tags\}|} \quad (1)$$

[22] http://www.treelogic.com/web/guest/soluciones/gestion-innovacion/-/journal_content/56/10137/12175

Recall (2) is the fraction of relevant tags that are successfully retrieved.

$$Recall = \frac{Match}{Count} = \frac{|\{Dataset\ Tags\} \cap \{Recommended\ Tags\}|}{|\{Relevant\ Tags\}|} \quad (2)$$

In both evaluation criteria we have employed the top-12 tags from the suggestions vector (precision@12 and recall@12). Following the approach in [24], it was assumed that if users do not find an accurate tag within the top-12 tags, they would skip the search and probably they would end up self-adding a new one.

6.2 Sample

We have employed a dataset comprised by a set of 483 tagged documents and 1049 tags. This dataset has been provided in the context of a research project by the partner Treelogic S.L. It is a slice of its research & development technological watching system and it has been generated by its workers in daily activities. These documents are both in English and Spanish, however there are a bigger number of documents in Spanish than in English. These information resources are in a variety of lengths and formats (e.g. text plain, HTML & XML) including typical data heterogeneities (e.g. special characters). The data was not pre-processed or cleaned trying to simulate the whole real scenario.

6.3 Results and Discussion

The precision and recall statistics are presented in **Table 4**, **Table 5** and **Table 6** . The statistics are divided into three columns. Firstly, precision@12 is located in the second column, while recall@12 in the third one, respectively.

A. Extraction Stage As it was expected, the results show that the combination of Normalization + NLP and Clustering obtains the highest precision and recall values. Since the test data is comprised by unigrams and n-grams and these techniques are focused on recommending unigrams and n-grams respectively. The whole set of tags cannot be matched by each separated technique, due to the composition of the test data.

Table 4. Extraction Stage precision@12 and recall@12

Heading level	Precision@12	Recall@12
A.1. Normalization	0.3737	0.0676
A.2. Normalization + NLP	0.3766	0.0682
A.3. Clustering	0.1144	0.0207
A. Extraction Stage	**0.4109**	**0.0744**

B. Enrichment Stage. The dataset contains a small percentage of English documents, as a consequence, querying to ontologies has almost no effect on the results. In contrast, querying to folksonomies increases dramatically both precision and recall.

Table 5. Enrichment Stage precision@12 and recall@12

Heading level	Precision@12	Recall@12
B.1. Ontology Querying	0.4156	0.0752
B.2. Folksonomy Querying	0.4290	0.0776
B. Enrichment Stage	**0.4290**	**0.0776**

C. Recommendation Stage. Starting from an empty database, in order to compute this value, as the satisfactory matches were found, they were used to feedback the system according to a chronological order with aim of simulating the real user's behaviour. Thus the outcome of this stage serves to increase the number of positive matches and therefore the accuracy of future recommendations avoiding the cold-start.

Table 6. Recommendation Stage precision@12 and recall@12

Heading level	Precision@12	Recall@12
C.1. Recommendation	0.4433	0.0802
C. Recommendation Stage	**0.4433**	**0.0802**

It is worth mentioning that the test data contain a certain proportion of tags which make reference to temporal events, internal projects or even personal matters. Therefore, the number of measurable tags are slightly reduced, and even a folksonomy beyond measure as Google Complete API is not able to suggest these unmatched tags.

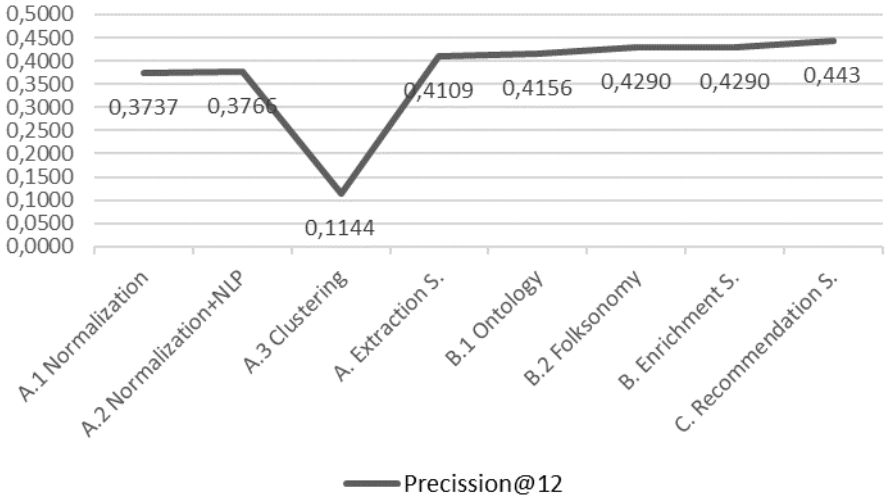


Fig. 4. Precision@12 Values

As it can be appreciated in both tables, recall@12 has low values. Due to the fact that the dataset-provided contains 1049 tags for 483 documents, the average of tags per documents is near 2.1718. Thus, the ideal recall@12 value would be 0.1810 (1049 tags tagged by experts per 5796 tags suggested by our system, 12 by each document). Therefore, the recall values are quite accurate, reaching values of 44.31%.

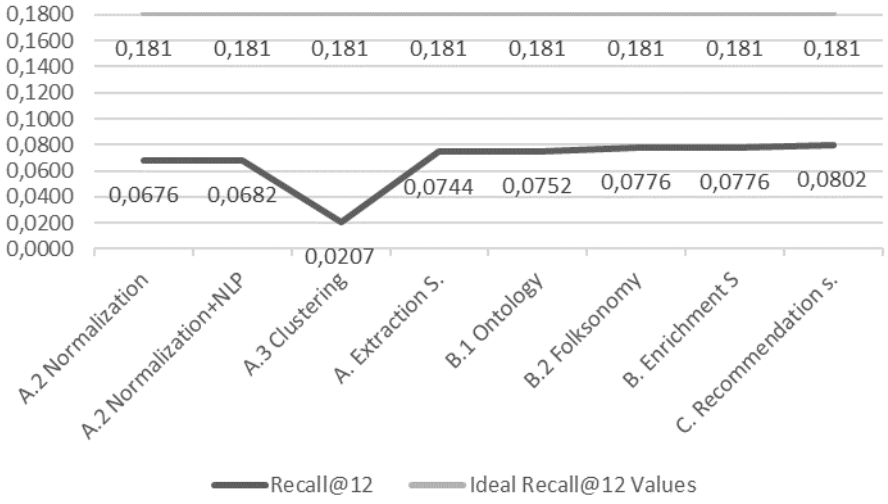


Fig. 5. Recall@12 Values

7 Conclusions and Future Work

We have proposed a methodology for tagging multilingual documents in a semiautomatic way, employing extraction, enrichment and user-behaviour recommendation techniques. Our methodology transfers the final decision of choosing the proper tag to the user, so if the recommended tags are non-suitable to him, he would add a new one giving feedback to the system. This feedback allows the recommendation engine to improve the results based on the previous behaviour of the users.

We have also developed a reference implementation, ACOTA, which was tested against production data from a research and development technology watching tool. Despite the considerable amount of internal and temporal tags, valid results have been obtained. This methodology and its implementation help knowledge workers to minimize the categorization-act effort providing a tool a better information resources classification within a knowledge organization.

As future work, we have considered increasing the amount of native languages, since English and Spanish are now supported; new languages can be added in order to internationalize the methodology and ACOTA as much as possible. As was aforementioned, tag compression is a technique which slightly boost the results, this feature can be partially supported (just in some ad-hoc languages, based on the requirements), helping to reduce the number of ambiguous tags in the system.

Furthermore the experimentation with new large and diverse datasets will give the proper feedback to improve the accuracy of each component. Another issue that we have taken into account is to improve the performance of the system. Although the system has a good performance with regular documents, it suffers when real-time (in terms of milliseconds) suggestions must be done processing large documents that contain several thousands of words. Finally a MapReduce[46] implementation of the proposed methodology has been considered, with the aim to categorize huge amounts of data such as digital libraries of governments or academic centres.

Acknowledgements

ACOTA is a subproject of the ORIGIN project, co-financed by ERDF (European Regional Development Fund), with the aim of increasing the productivity of software development activities in global scenarios.

References

1. Belkin, N., Croft, W.: Information filtering and information retrieval: two sides of the same coin? Communications of the ACM. 29, 1–10 (1992).
2. Shklar, L., Sheth, A., Kashyap, V., Shah, K.: InfoHarness: Use of automatically generated metadata for search and retrieval of heterogeneous information. Advanced Information Systems Engineering. pp. 217–230 (1995).
3. Large, A., Moukdad, H.: Multilingual access to web resources: an overview. Program: electronic library and information systems. 34, 43–58 (2000).
4. W3Techs: Usage of content languages for websites, http://w3techs.com/technologies/overview/content_language/all.
5. Kunder, M. de: The size of the World Wide Web (The Internet), http://www.worldwidewebsize.com/.
6. Hjørland, B.: Semantics and knowledge organization. Annual Review of Information Science and Technology. 41, 367–405 (2007).
7. Hjørlandd, B.: Methods for evaluating information sources: An annotated catalogue. Journal of Information Science. 38, 258–268 (2012).
8. Gruber, T.: Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies. 43, 907–928 (1995).
9. Wal, T. Vander: Folksonomy Coinage and Definition, http://vanderwal.net/folksonomy.html.
10. Casado-Lumbreras, C., Rodríguez-González, A., Álvarez-Rodríguez, J.M., Colomo-Palacios, R.: PsyDis: Towards a diagnosis support system for psychological disorders. Expert Systems with Applications. 39, 11391–11403 (2012).
11. García-Crespo, Á., Rodríguez, A., Mencke, M., Gómez-Berbís, J.M., Colomo-Palacios, R.: ODDIN: Ontology-driven differential diagnosis based on logical inference and probabilistic refinements. Expert Systems with Applications. 37, 2621–2628 (2010).
12. Villazón-Terrazas, B., Ramírez, J., Suárez-Figueroa, M.C., Gómez-Pérez, A.: A network of ontology networks for building e-employment advanced systems. Expert Systems with Applications. 38, 13612–13624 (2011).
13. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. The Semantic Web: Research and Applications. 4011, 411–426 (2006).
14. Yoo, D., Choi, K., Suh, Y., Kim, G.: Building and evaluating a collaboratively built structured folksonomy. Journal of Information Science. (2013).

15. Córdoba, a., Astrain, J.J., Villadangos, J., Echarte, F.: A self-adapted method for the categorization of social resources. Expert Systems with Applications. 40, 3696–3714 (2013).
16. Tourné, N., Godoy, D.: Evaluating tag filtering techniques for web resource classification in folksonomies. Expert Systems with Applications. 39, 9723–9729 (2012).
17. Carrer-Neto, W., Hernández-Alcaraz, M.L., Valencia-García, R., García-Sánchez, F.: Social knowledge-based recommender system. Application to the movies domain. Expert Systems with Applications. 39, 10990–11000 (2012).
18. Kiu, C.-C., Tsui, E.: TaxoFolk: A hybrid taxonomy–folksonomy structure for knowledge classification and navigation. Expert Systems with Applications. 38, 6049–6058 (2011).
19. Shirky, C.: Ontology is Overrated: Categories, Links, and Tags, http://www.shirky.com/writings/ontology_overrated.html?goback=.gde_1838701_member_179729766.
20. Halpin, H., Robu, V., Shepard, H.: The dynamics and semantics of collaborative tagging. Proceedings of the 1st Semantic Authoring and Annotation Workshop SAAW'06. (2006).
21. Newman, M.: Power laws, Pareto distributions and Zipf's law. Contemporary physics. (2005).
22. Park, S.-T., Pennock, D., Madani, O., Good, N., DeCoste, D.: Naïve filterbots for robust cold-start recommendations. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06. pp. 699–705. ACM Press, New York, New York, USA (2006).
23. Brook, C.H., Montanez, N.: Improved annotation of the blogopshere via autotagging and hierarchical clustering. Proceedings of the 15th World Wide Web Conference (WWW06). (2006).
24. Mishne, G.: AutoTag. Proceedings of the 15th international conference on World Wide Web (WWW 06). p. 953. ACM Press, New York, New York, USA (2006).
25. Sood, S.C., Owsley, S.H., Hammond, K.J., Birnbaum, L.: TagAssist: Automatic Tag Suggestion for Blog Posts. (2007).
26. Noll, M.G., Au Yeung, C., Gibbins, N., Meinel, C., Shadbolt, N.: Telling experts from spammers. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09. p. 612. ACM Press, New York, New York, USA (2009).
27. Hong, L., Ahmed, A., Gurumurthy, S.: Discovering geographical topics in the twitter stream. Proceedings of the 21st international conference on World Wide Web. 769–778 (2012).
28. Gayo-Avello, D., Álvarez-Gutiérrez, D., Gayo-Avello, J.: Naïve Algorithms for Keyphrase Extraction and Text Summarization from a Single Document Inspired by the Protein Biosynthesis Process. Biologically Inspired Approaches to Advanced Information Technology. LNCS 3141, 440–455 (2004).
29. Mika, P., Ciaramita, M., Zaragoza, H., Atserias, J.: Learning to Tag and Tagging to Learn: A Case Study on Wikipedia. IEEE Intelligent Systems. 23, 26–33 (2008).
30. Song, Y., Zhuang, Z., Li, H., Zhao, Q., Li, J., Lee, W.-C., Giles, C.L.: Real-time automatic tag recommendation. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08. 515 (2008).
31. Sun, K., Wang, X., Sun, C., Lin, L.: A language model approach for tag recommendation. Expert Systems with Applications. 38, 1575–1582 (2011).
32. Dostal, M., Ježek, K.: Automatic keyphrase Extraction extraction based on NLP Automatic Keyphrase based on NLP and statistical methods and Statistical Methods. Proceedings of the Dateso 2011: Annual International Workshop on DAtabases, TExts, Specifications and Object. 140–145 (2011).
33. Labra Gayo, J.E., De Pablos, P.O., Cueva Lovelle, J.M.: WESONet: Applying semantic web technologies and collaborative tagging to multimedia web information systems. Computers in Human Behavior. 26, 205–209 (2010).
34. Jimenez-Nácero, W., Luis-Alvargonzález, C., Abella-Vallina, P., Alvarez-Rodríguez, J.M., Labra-Gayo, J.E., Ordoñez de Pablos, P.: Emergent Ontologies by collaborative tagging for Knowledge Management. Advancing Information Management through Semantic Web Concepts and Ontologies. p. 16. IGI-Global (2012).
35. Kern, R., Granitzer, M., Pammer, V.: Extending Folksonomies for Image Tagging. 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services. pp. 126–129. IEEE (2008).
36. Chen, P.-I., Lin, S.-J.: Automatic keyword prediction using Google similarity distance. Expert Systems with Applications. 37, 1928–1938 (2010).
37. Sigurd, B., Eeg-Olofsson, M., Van Weijer, J.: Word length, sentence length and frequency - Zipf revisited. Studia Linguistica. 58, 37–52 (2004).
38. Suchanek, F.M., Vojnovic, M., Gunawardena, D.: Social tags. Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08. p. 223. ACM Press, New York, New York, USA (2008).
39. Miller, G. a.: WordNet: a lexical database for English. Communications of the ACM. 38, 39–41 (1995).
40. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. Proceedings of the tenth international conference on World Wide Web - WWW '01. pp. 285–295. ACM Press, New York, New York, USA (2001).
41. Lipkus, A.: A proof of the triangle inequality for the Tanimoto distance. Journal of Mathematical Chemistry. 26, 263–265 (1999).
42. Tanimoto, T.T.: An elementary mathematical theory of classification and prediction. (1958).

43. Davidson, C.H.: Technology watch in the construction sector: why and how? Building Research & Information. 29, 233–241 (2001).
44. Keats, J.: Virtual Words: Language on the Edge of Science and Technology. Oxford University Press, Inc., Oxford, New York (2010).
45. Cleverdon, C., Mills, J., Keen, M.: Factors determining the performance of indexing systems. (1966).
46. Dean, J., Ghemawat, S.: MapReduce. Communications of the ACM. 51, 107 (2008).