Webcrawling clustering en espacio multidimensional basado en distancia y su aplicación a Opinion Mining (•••)

Ezequiel Gorbatik•, Hugo O. Barrera•, E. Schneider Loaiza•,
Fabián Riaño Santiesteban•, Francisco Gindre.•
M. Daniela López De Luise. ••

Resumen

La explosión multimedial y la revolución surgida a partir de la Web 2.0 donde los consumidores de información son a su vez productores de contenido han reflejado un cambio de paradigma en la comunicación.

Este cambio vuelve a las herramientas de sondeo de opinión tales como encuestas, focus group y sondeos telefónicos limitadas en su alcance, imprecisas en sus resultados y sesgadas por sus métodos convirtiendo a las mismas en prácticamente obsoletas.

Los medios de comunicación se han hecho eco de esto y en los medios más representativos se permite a los lectores participar de las noticias por medio de las redes sociales.

Son necesarias para explorar y descubrir de manera continua y sistematizada estos nuevos canales de expresión, analizar los contenidos de las expresiones en los mismos y poder extraer conocimiento de estos flujos de información a escala masiva.

En este trabajo se parte de un nuevo concepto de la minería de datos, se analiza una nueva estrategia para descubrir nuevos canales mediante el Webcrawling inteligente, se proponen nuevos formas de modelado de conceptos y opiniones para poder sintetizarlos y cuantificarlos para su posterior análisis, se da a conocer un método para realizar este análisis de las percepciones y finalmente se demuestran las posibilidades de clusterización de la información obtenida.

Palabras Clave: minería de datos, Clustering, minería de opinión, Webcrawling

Fecha de recepción: agosto 2012 | Fecha de aceptación: octubre 2012

• Investigadores del AIGroup, Universidad de Palermo

•• Directora del AIGroup, Universidad de Palermo

••• Este trabajo fue aprobado y presentado en el Congreso TRIC V organizado por IEEE CIS (Computational Intelligence Society) Argentina que se desarrolló en el marco del IEEE Argencon 2012 del 13 al 15 de junio de 2012 en la UNC.

Abstract

Multimedia consumption and the revolution caused by the Web 2.0 phenomenon, where information consumers are also producers are clearly reflected by a paradigm shift in modern communication. Due to this change, traditional survey tools such as focus groups, phone surveys and quizzes are now limited in scope and lack precision, turning them obsolete. Mass media has echoed this and now allows readers and consumers to participate in news through social networks.
This paper analyzes a new strategy to discover new channels, proposes new ways to model concept and opinions and how to synthetize and quantify them for their later analysis, and, finally, various methods for clustering and grouping for the obtained information are shown.

Keywords: Data mining, Fuzzy Clustering, Opinion mining, Text mining, Webcrawling

I. Introducción

Un sistema es un conjunto de componentes interrelacionados que recibe información de entrada, la procesa y responde una salida [28].

Ésta definición es aplicable a fenómenos físicos, personas y sistemas de información.

Un sistema informático en general tiene entradas, procesos y salidas determinadas. Dentro de los sistemas mencionados tenemos aquellos que se especializan en la adquisición de datos (Crawlers, webBots), procesamiento (Software de procesamiento estadístico), almacenamiento (Sistema de Gestión de Bases de Datos) y devolución de datos (Generadores de Reportes). Estos sistemas en conjunto forman parte de la minería de datos.

Los sistemas de soporte de decisión llevan el procesado de datos a otro nivel, extrayendo información a partir de inferencias estadísticas avanzadas aplicadas a los mismos.

A partir de las capacidades de análisis de los sistemas de minería de datos y soporte de decisión combinándolos con herramientas de análisis de texto se llega a la minería de texto que consiste en la extracción de información a partir de textos, no sólo de datos tabulados.

Utilizando herramientas avanzadas de análisis de texto y procesamiento de lenguaje natural se pueden hacer inferencias sobre las opiniones acerca de un concepto determinado [17].

En este trabajo se analiza la minería de datos como el pasaje de la información y los datos por distintos estados de agregación.

Partiendo de la información textual, los datos tabulados, la información procesada y los "puntos de vista" descubiertos se pueden extraer parámetros estadísticos acerca de las percepciones sobre un determinado concepto censando distintas fuentes de información haciendo minería de opinión.

II. Minería de Datos

"La minería de datos es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de los datos" [16].

Sintetizando la definición anterior el fin último de la minería de datos es la extracción de información a partir de los mismos.

Dentro de ésta podemos destacar algunos elementos tangibles como será el volumen de los datos de entrada en sí, un conjunto de algoritmos de procesamiento, el análisis estadístico y finalmente las técnicas de visualización de la información obtenida.

Además de la efectividad de los algoritmos y la calidad de la representación; los datos en sí poseen atributos que determinaran los resultados. Estos atributos son precisión de datos, nivel de corrección de los mismos, nivel de actualización, completitud, integridad y relevancia.

Más allá de ésta definición tradicional de minería de datos para avanzar un poco más en los conceptos presentados en este documento se propone utilizar una definición alternativa que define la minería de datos como el resultado de la interacción dinámica de distintos estados de agregación de la información [22].

Esta definición indica que la minería de datos consiste en el pasaje de los datos de un estado de agregación inicial "D" que contiene un conjunto de datos puros sin procesar en otro estado de agregación final que permita al usuario obtener un beneficio de este proceso.

Se identifican cuatro estados de agregación. El primer estado es el "D" de datos puros sin procesar, en este caso importan además de los atributos mencionados el volumen y el origen.

Al analizar estos datos con métodos estadísticos surge información que puede ser de algún interés tales como la media, la mediana, fractiles, etc. Esta información (en esta etapa ya no son datos) está en un segundo estado de agregación "F" en este estado se empieza a determinar si los datos poseen algún significado.

A partir de los datos "D" y la información "F" podemos comenzar a realizar un análisis más detallado, depurando los datos, segmentándolos y sacando conclusiones "puntos de vista" sobre los mismos; este es el estado "I".

La información puede fluir entre los distintos estados de agregación también en sentido inverso por ejemplo para verificar las conclusiones del estado "I" contrastándolas con las obtenidas en el estado "F". También la información puede ir del estado "I" al "D" al verificar las hipótesis con los datos originales más allá de los indicadores.

Existe un cuarto estado "K" que es el conocimiento que resulta de las conclusiones obtenidas a largo plazo; las proyecciones y tendencias es información en este estado.

III. Origen de datos: WebCrawling Inteligente

El primer problema con el que se debe lidiar es con el origen de los datos para el cual se necesita una exploración inteligente de los sitios, tomando en cuenta que se está realizando un sondeo de opinión que en cierta medida partirá de la sintetización, cuantificación y clasificación de conceptos es razonable realizar el *Crawling* de los datos bajo los mismos principios; paralelamente a este trabajo se está trabajando en un WebCrawler inteligente que lidia con estos problemas.

A. Proceso de Crawling Inteligente

El proceso de *Crawling* se basa en una arquitectura de agentes [25] y tiene 3 pasos básicos que corren en secuencia para una misma web pero en paralelo para el conjunto. Se parte de una lista inicial, se descarga el contenido (En el proyecto se utilizaron las librería BoilerPipe [14] y JSoup [15]) y se extraen los links para agregarlos a la lista. Durante el proceso de C*rawling* la lista de links crece de manera exponencial con cada exploración, lo cual representa un problema en términos de la dicha complejidad computacional. Para solventar esto debe hacerse una selección óptima de links de acuerdo a su relevancia; un filtrado y una priorización inteligente.

Como se mencionó anteriormente, este trabajo propone la búsqueda y organización sistematizada a partir de la síntesis y cuantificación de los significados y conceptos para que a partir de la misma se pueda obtener una ponderación óptima para cada link.

Considerando la síntesis de un concepto como la unión de otros términos y conceptos, el primer problema a resolver es como expresar esta relación, cómo medirla y qué propiedades tiene. Se debe tener una distancia entre términos, luego se debe tener un espacio multidimensional donde expresar estas distancias, también se debe delimitar que se considera parte del concepto y que no; esto a su vez dependerá de la retroalimentación de nueva información, el poder expresar nuevos documentos "aprendidos" en función de las distancias hacia los conceptos a buscar, para finalmente poder jerarquizar los vínculos en función de la información procesada. En las siguientes secciones serán tratados todos estos puntos.

1) Relaciones entre Términos

Como se menciona en los párrafos anteriores el primer problema a resolver es cómo expresar la relación entre distintos términos para expresar un concepto.

Durante el curso de la presente investigación se evaluaron las propiedades que tendría que tener esta relación. La característica principal es que debe expresar en una medida aproximadamente uniforme que precise cuán "cerca" está un término de un concepto. Esta aproximación no tiene que ser precisa (la distancia de un concepto a otro es "difusa") sin embargo debe expresar un "orden" de cercanía. Un término debe estar más relacionado que otro en referencia a un concepto.

Para continuar con la explicación se deben hacer las siguientes definiciones

- **Palabra:** Unidad sintáctica mínima e indivisible con significado.
- **Frase:** Dos o más palabras agrupadas cuyo significado es distinto del significado literal de las mismas.

• **Término:** Palabra o frase que representa total o parcialmente un concepto.

• **Concepto:** Entidad formada por un conjunto difuso de términos y relaciones que engloban toda la información relevante y conocida de la entidad.

2) Conjuntos y Funciones

Partiendo de las definiciones anteriores se deben definir los conjuntos (Ver Tabla I) y funciones asociados (Ver Tabla II,III y IV). Cabe destacar que el espacio donde se expresarán los documentos es un espacio vectorial [3] en principio en $\Re^3$ pero pueden haber tantas dimensiones como se necesiten.

Los ejes de este espacio pueden tener distintas escalas, en los casos presentados en este documento una escala es logarítmica y continua mientras que la otra es lineal y discreta.

Tabla I: Conjuntos

$P = \{Conjunto\ de\ Palabras\}$
$x/x \in P : x\ es\ "Palabra"$
$F = \{Conjunto\ de\ Frases\}$
$x/x \in F : x = a \oplus b \oplus .. \oplus n\ ; a, b, .., n \in P$
$T = \{Conjunto\ de\ Términos\}$
$T = P \cup F$
$S = \{Relación\ de\ distancias\ entre\ Términos\}$
$s : T \times T \rightarrow\ (0 \leq s \leq 1) \subseteq \mathbb{R}$
$S = \{(x, y, z) : x, y \in T \wedge z \in \mathbb{R}\ /\ s(x, y) = z\}$
$s : T \times T \rightarrow \mathbb{R}\ /\ s = Distancia\ Semántica\ entre\ x\ e\ y$
$C = \{Conjunto\ de\ Conceptos\}$
$C = T \cup S$

Los conjuntos más relevantes son T, S y C

Tabla II: Axiomas de la función S

$s(x, x) = 0$
$s(x, \bar{x}) = \infty$

Tabla III: Límites de función S

$\lim_{y \to x} s(x, y) = 0$
$\lim_{y \to \bar{x}} s(x, y) = \infty$

Tabla IV: Propiedades de la función S

Reflexividad	$\forall x \in T, \quad s(x,x) = 0 \Rightarrow \forall x \ni s(x,x)$ La relación de un concepto consigo mismo es "0".
Simetría	$\forall x, y \in T, \quad s(x,y) \cong s(y,x)$ La relación del concepto "a" con el "b" será similar a la del b con a
Intransitiva	La función "s" es intransitiva pero la transitividad se resuelve ponderando los saltos.
No antisimétrica	$\forall x, y, z \in T, s(x,y) = s(z,y) \Rightarrow x = z$ La igualdad de la distancia entre conceptos no implica la igualdad de conceptos.

Las opciones para esta función surgidas de la investigación son las siguientes:
LSA[1], PMI[2][7], VGEM[3], NGD (Distancia Google Normalizada) [4][5] y SimRank[6]. Por razones de espacio se explicará como ejemplo NGD.

3) Distancia Google Normalizada

La distancia normalizada de Google (NGD) es una medida de similitud semántica obtenida a partir de los resultados del motor de búsqueda de Google para un conjunto determinado de *Keywords*.[4][5]

Aquellos *Keywords* altamente relacionados tenderán a valores más cercanos entre sí que las que tienen significados distintos.

La función de distancia semántica de NGD en está definida de la siguiente manera:

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}} \qquad (1)$$

Ecuación 1: Distancia Google Normalizada

Donde M es el número total de páginas indexadas en Google; $f(x)$ y $f(y)$ son el número de Hits de los términos "x" e "y"; $f(x, y)$ son el número de veces que ambos términos aparecen en simultáneo. [8]

La ecuación cumple con las propiedades y los axiomas anteriormente mencionados, si dos términos no aparecen en simultáneo pero aparecen separados la medida tenderá a infinito, si ocurren siempre de manera simultánea la ecuación tenderá a cero.

Nótese que es una escala logarítmica y que existen algunos sesgos que deben tomarse en cuenta tales como las palabras homónimas (que mezclarían nubes de puntos distintas) y sinónimos (que superpondrían varias nubes de puntos, con puntos

cercanos a 0); también existen sesgos aleatorios de palabras que muestran conceptos no relacionados como si estuvieran altamente relacionados. Tales sesgos culturales de palabras que en una determinada cultura están altamente relacionados por una "excepción" cultural pero que conceptualmente no están relacionados (Ejemplo: en Argentina, "pumas" y "rugby", "leonas" y "hockey").

Si bien se menciona como distancia Google porque refiere a los resultados de este buscador en particular lo correcto sería hablar de una medida en base a la distribución de frecuencias de las presencias de términos en textos conocidos.

4) Ejemplo de Cuantificación y Sintetización

A continuación se da un ejemplo del sistema propuesto (Ver Tabla V, VI, VII, VIII).

Se utilizarán los términos *"Marsupial", "Guerra", "Paz", "Armas"* y *"Violencia"*. Claramente se puede intuir que *"Marsupial"* es distinto que los demás términos.

A continuación se demuestra cómo se cuantificará esta diferencia intuitiva ya que la exploración inteligente deberá hacer una interpretación de las páginas web exploradas a fin de poder ordenar tanto los links que las contienen como los links que las apuntan.

El objetivo de ésta cuantificación será tener un conjunto de puntos en un plano que expresen las relaciones entre los términos y consecuentemente muestren una clara diferencia entre el término *"Marsupial"* y los otros términos.

En principio se obtienen los *Hits* (Cantidad de resultados de Google) de los términos.

Se obtienen los *Hits* individuales (en la tabla es la intersección de un término consigo mismo) y los *Hits* grupales (intersección con otros términos en la tabla). Los *Hits* grupales implican la presencia de ambos términos en una página Web en cualquier orden.

Tabla V: Google Hits

Keyword /F(X)	Marsupial	Guerra	Paz	Armas	Violencia
Marsupial	**3.680**	892	814	234	435
Guerra	892	**419**	88.8	56.2	174
Paz	814	88.8	**507**	54.9	64.2
Armas	234	56.2	54.9	**121**	36.3
Violencia	435	174	64.2	36.3	**143**

La unidad es miles de hits, las búsquedas se hicieron el día 21 de diciembre del 2011.

Luego se procede con el cálculo de las distancias normalizadas con la ecuación antes mencionada.

Tabla VI: Distancia Google Normalizada

Keyword /NGD	Marsupial	Guerra	Paz	Armas	Violencia
Marsupial	**0**	0,4153	0,4343	0,4217	0,3912
Guerra	0,4153	**0**	0,1073	0,0472	-0,0121
Paz	0,4343	0,1073	**0**	0,0484	0,0491
Armas	0,4217	0,0472	0,0484	**0**	0,0780
Violencia	0,3912	-0,0121	0,0491	0,0780	**0**

Como se desea ver el contraste entre las distancias desde y hacia el término *"Marsupial"*, se analizará la columna correspondiente siendo que ésta será utilizada para formar el centroide del clúster.

Tabla VII: Distancia Google Normalizada al centroide Marsupial

	Marsupial
Marsupial	**0**
Guerra	0,4153
Paz	0,4343
Violencia	0,4217
Armas	0,3912

A continuación se presenta un gráfico, el eje X en realidad se usa para separar los conceptos, no tiene una medida y el eje Y contiene las distancias previamente calculadas. Debe considerarse que los puntos están en la misma línea vertical y que el eje Y tiene una escala logarítmica.

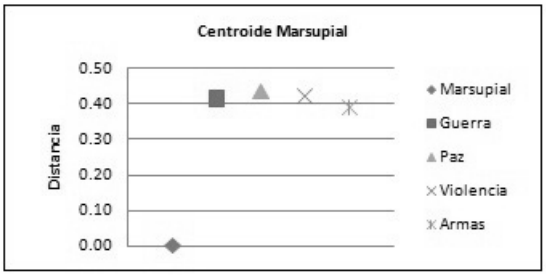

Gráfico 1: Distancias hacia el centroide Marsupial basado en NGD

Para poder analizar las distancias desde otro centroide se analiza a continuación el centroide *"Paz"*.

Tabla VIII: Distancia Google Normalizada al centroide Paz

	Paz
Marsupial	0,4343
Guerra	0,1073
Paz	0
Violencia	0,0484
Armas	0,0491

Se puede observar que salvo el concepto "Marsupial" todos los conceptos se "acercan" a paz, lo cual es consistente con el cluster anterior.

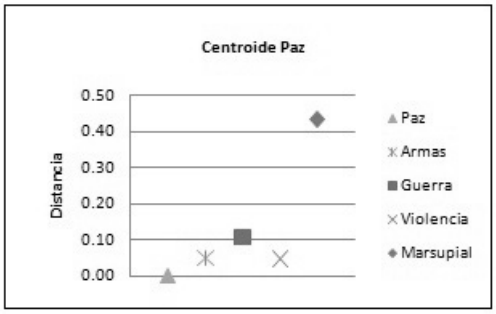

Gráfico 2: Distancias hacia el centroide Paz basado en NGD.

Estas distancias ilustran la relación entre conceptos. Por ejemplo, la guerra y la paz son antónimos, si bien los significados de estos conceptos son opuestos, está relación (como cualquier relación de antónimos) es una relación fuerte. Un concepto se define tanto por lo que es, como por lo que no es y estas relaciones están expresadas como se muestra en el ejemplo.

Estas nubes de puntos forman áreas de mayor densidad de puntos que otras, analizando esas densidades en relación a un radio puede definirse el límite del clúster para buscar puntos de los cuales se tomarán los links.

a) Distancia Levenshtein

La distancia Levenshtein [9] es una distancia entre cadenas de caracteres y consiste en el número mínimo de operaciones para transformar una palabra

en otra, con la distancia Levenshtein se puede obtener una medida de palabras asociadas entre sí generadas o asociadas a partir de algoritmos de stemming [10] [11][12][13].

Esta distancia es escalar, no implica un acercamiento conceptual (Por ejemplo Paz y Pez tienen una distancia de una unidad pero no son conceptos relacionados conceptualmente).

Tabla IX: Distancia Levenshtein a palabras asociadas a "Guerra"

	Levenshtein
Guerra	0
Guerras	1
Guerrear	2
Guerrero	4

Tabla X: Distancia Levenshtein a palabras asociadas a "Marsupial"

	Levenshtein
Marsupial	0
Marsupiales	2
Marsopa	4

b) Espacio

A partir de las distancias (NGD) obtenidas combinándolas a su vez con las distancias de cadenas (Levenshtein) anteriormente mencionadas en un espacio tridimensional podemos crear una serie de puntos asociados entre sí ya sea por significados, por distancia de cadena y/o por cualquier combinación de ellos.

En principio este espacio es básicamente tridimensional, pero si tomamos en cuenta idioma, tipo de palabra, dominio y/o presentamos cada palabra como una superposición de planos estaríamos expresando los términos y conceptos en un espacio multidimensional, siendo la relación expresada en la distancia espacial pudiéndose obtener una distancia absoluta entre términos a partir del modulo de los vectores asociados.

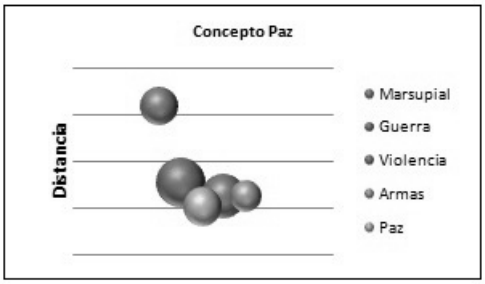

Gráfico 3: Gráfico tridimensional formado a partir de distancias obtenidas.

c) Clustering Difuso

A partir de la información obtenida expresada en el espacio los conceptos estarán expresados como nubes de puntos en el espacio multidimensional [18][19], si consideramos a los términos buscados como los centroides de un clúster [20] podemos definir una radio arbitrario para determinar qué puntos entran dentro del concepto y que puntos no [21], el radio dependerá del concepto a buscar, la densidad de la nube de puntos, el tipo de palabra del keyword, etc.

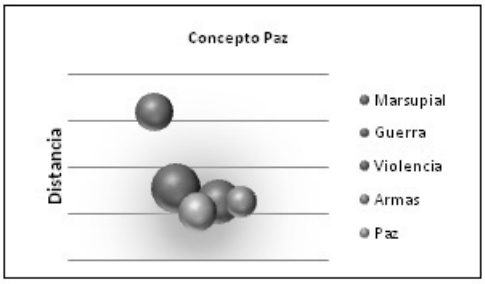

Gráfico 4: Gráfico tridimensional formado a partir de distancias obtenidas con una proyección de la nube difusa abarcando los puntos cercanos.

d) Jerarquización de Links

Considerando que los puntos estarán asociados a las URL de origen que los suscribieron, una vez definido el radio multidimensional difuso, puede usarse un algoritmo de consenso para la jerarquización de links.

Suponiendo que se buscan páginas sobre la Paz, nuestro universo de términos está compuesto por *"Marsupial", "Guerra", "Violencia", "Armas"* y *"Paz"*. Nuestro clúster de "Paz" dejó excluyó al término *"Marsupial"* pero incluyo al resto.

El proceso es el siguiente: Se relevan *n* páginas, de esas páginas relevadas, hay una página que incluye solo *"Paz"*, otra página *"Paz"* y *"Armas"* y otra que incluye todos los términos de nuestro universo excepto "*Marsupial"*. Cada punto de nuestra nube de puntos tendrá un listado de links de páginas que los contienen. Al relevarse todos los puntos de un área densa habrán algunos sitios que estarán en más listados y en función de eso se infiere la relevancia, el orden y la prioridad para seguir explorando.

B. Articulación con sistema de opinion mining

A partir de los resultados de la exploración se extraen las oraciones (Mediante OpenNLP [23]) que serán el punto de entrada del sistema de *Opinion Mining* (Proyecto JAZZ)

La extracción de texto y segmentación de oraciones tiene como finalidad generar archivos de texto plano con la información textual contenida en los resultados de la búsqueda ingresada. Los sitios están agrupados por dominio, utilizando una estructura de carpetas con nomenclatura de dominio reverso (ej: ar.com.aigroup) a fin de identificar a qué sitio pertenecen los archivos de texto.

El proceso de extracción separa el cuerpo de texto encontrado en un sitio en oraciones. Estas oraciones son guardadas cada una en una línea diferente en el archivo de texto que representa ese archivo HTML

IV. Minería de Opinión: Keywords, Pesos y Frecuencias

Una vez obtenidos los datos se comienza con la minería de opinión en sí. En principio se cuentan con un gran volumen de información en estado D, que han sido previamente depuradas en su origen en relación al concepto objetivo pasando al estado de agregación F.

Se procede a analizar una oración, se asume que mantiene cierta relación con el concepto objetivo ya que ha sido obtenida a partir del Webcrawling inteligente, no obstante para reforzar más los resultados se buscará el concepto central y conceptos asociados (obtenidos del webcrawler) para reforzar o debilitar la opinión hallada en la oración.

La ecuación para determinar la opinión (Función Op) de un concepto (c) en una oración (o) se determina a partir la siguiente fórmula:

$$Op(c,o) = (Bc \quad Nc \quad Mc)Fc \qquad (2)$$

Ecuación 2: Función Opinión

La función Op (Ecuación 2) tiene como entrada un concepto y una oración y devuelve 3 elementos que identificarán que tan Positivo, Neutral o Negativo (Bc, Nc, Mc) es "opinado" ese concepto en el contexto de esa oración (o) ponderado por la fuerza (Fc) de la presencia de ese concepto en la misma.

Tabla XI: Valores del Factor Fc (Factor de refuerzo)

Concepto en oración	Factor de Refuerzo
Ausente	0,5
Mismo Concepto	1
Concepto asociado	Complemento de la Distancia Semántica Google (Obtenida del Webcrawler)

Por ejemplo en la siguiente oración se busca analizar el concepto "Messi":

"Messi es el mejor jugador del mundo."

Intuitivamente sabemos que en la oración hay una opinión positiva del concepto "Messi", como mencionamos anteriormente primero determinamos la presencia del concepto, de acuerdo a la tabla anterior (Tabla XI) tiene una fuerza (Fc) de 1.

Si la oración dijera *"El delantero del Barcelona es el mejor del mundo."* el factor *Fc* baja al 61% (Resultados obtenidos el 25/3/2012), se aprecia que si bien el significado es el mismo en el primer caso la apreciación es *personal* y *directa* pero en el segundo caso no.

Y si la oración dijera simplemente *"Es el mejor del mundo."* la fuerza de la opinión se reduce a un 50% ya que no es personal, no hay mención directa ni indirecta aunque se infiere que está relacionado con el concepto a partir de la resolución del Webcrawler.

El punto más interesante de la minería de opinión es como se determina una opinión positiva, neutral o negativa.

En este caso se parte de pesos, frecuencias y relaciones de keywords.

En el ejemplo anterior se entiende que la palabra *"mejor"* indica claramente una opinión positiva.

Si en cambio dijera *"Messi es un grande."* la palabra *"grande"* también indica una opinión positiva, si en vez de grande dijera "pésimo" se entendería claramente una opinión negativa esto se ilustra en la siguiente tabla (Tabla XII).

Tabla XII: Clasificadores de Opinión

Positivo	Neutral	Negativo
Bueno	Mediocre	Mal
Lindo	Promedio	Terrible
Impresionante	Suficiente	Defectuoso
Excelente	Limpio	Pésimo
Grande	Bien	Pobre

Esto traducido en información para la ecuación antes mencionada queda de la siguiente manera (Tabla XIII).

Tabla XIII: Clasificadores de Opinión

Keyword	B	N	M
Bueno	1	0	0
Lindo	1	0	0
Impresionante	1	0	0
Excelente	1	0	0
Grande	1	0	0
Mediocre	0	1	0
Promedio	0	1	0
Suficiente	0	1	0
Limpio	0	1	0
Bien	0	1	0
Mal	0	0	1
Terrible	0	0	1
Defectuoso	0	0	1
Pésimo	0	0	1
Pobre	0	0	1

Por razones de espacio la tabla ha sido acotada, sin embargo existen términos que no son ni absolutamente positivos, ni absolutamente negativos, ni absolutamente neutrales ya sea por el criterio de búsqueda o que simplemente no figuran en la base de conocimiento del motor de opinión.

Para estos casos hay que buscar similitudes entre las palabras encontradas y las conocidas.

En el proyecto GDARIM (WebCrawler inteligente) se utiliza una escala para medir similitudes entre conceptos que si bien es una medida aceptable para medir

distancias entre los clasificadores conocidos y desconocidos, existen alternativas de búsqueda con resultados más precisos como se muestra a continuación.

Tabla XIV: Distancias de Percepción de Conceptos Desconocidos

Palabra	B	N	M	Observación
Bueno	1			Conocido
Hermoso	1			Sinónimo: "Lindo"
Admirable	0,25			Sustantivo: "Admiración"
				Sinónimo: "Asombro"
				Adjetivo: "Asombroso"
				Sinónimo: "Bueno"
Capaz		0,5		Sinónimo: "Apto"
				Sinónimo: "Suficiente"

El análisis de realizó utilizando información de Wikitionary en español (Datos tomados 25/03/2012), las reglas de ponderación son las siguientes:

Se inicia de puntaje ideal "1", sinónimo directo a concepto conocido se mantiene sinónimo de sinónimo se divide por 2 el puntaje actual; pasaje a sustantivo se vuelve a dividir por 2 y así sucesivamente hasta encontrar un calificador conocido o el valor baje de 0,05.

Cuanto mayor sea la cantidad de calificadores conocidos mejor será la precisión de las opiniones detectadas.

Lo más importante del proceso anterior es que éste es replicable ya que con el mismo set de datos y con el mismo grupo calificadores conocidos se obtiene el mismo resultado y una medida uniforme para buscar opiniones sobre conceptos diversos.

Por otro lado esta estrategia permite el modelado de la escala valor de manera subjetiva permitiendo ponderar arbitrariamente conceptos específicos y medir distancias de percepción (Tabla XIV) a partir de esos valores.

V. Trabajo a futuro: Clustering de Opiniones

De la misma forma que se hizo *clustering* de conceptos en el proceso de crawling se puede hacer *clustering* de las opiniones ya sea por percepciones positivas, neutrales y negativas; en relación a calificadores específicos o incluso también en referencias temporales (si se toma en cuenta el tiempo de los datos de entrada).

Contando con un set de datos lo suficientemente grande se pueden buscar patrones de opinión ya que se contará con set de datos abundantes, información de contexto suficiente (lugar de origen, fecha, sitio, etc.) e incluso una escala ordenada

de valor de las palabras claves asociadas a las percepciones que permitirán hacer *clustering* y buscar patrones respecto a clasificadores específicos.

Por ejemplo si se busca hacer un sondeo de opinión de un artículo electrónico y se halla una alta concentración de puntos entorno a los conceptos "rápido" y "buen diseño" es posible realizar un análisis sobre esos aspectos; pero si se analizan más profundamente los resultados mediante algún filtrado de los datos (por ejemplo tomando en cuenta los países de donde originaron las oraciones) podrían apreciarse diferencias de las cuales se podría inferir que en general todos los países consideran un producto como "rápido" pero hay diferencias en cuanto a la apreciación de la estética del producto respecto de cada país.

La estrategia utilizada para modelar las fuentes de datos, los conceptos a sondear y a descubrir, las distancias entre conceptos y el modelado de la percepción hacen posible prácticamente cualquier tipo de segmentación geográfica (país de origen de los datos), demográfica (a partir de datos demográficos de la fuente de datos) e idiomática (idioma de origen).

VI. Conclusiones

En el presente trabajo se muestra el marco teórico para el modelado de conceptos y percepciones así como estrategias para la sintetización y cuantificación de los mismos.

Es posible cuantificar la distancia entre conceptos y esta medida representa mejoras sustanciales para la búsqueda de nuevas fuentes de información de manera óptima y automatizada.

La cuantificación y sintetización de las percepciones con una medida uniforme permite hacer sondeos estadísticos sobre grandes conjuntos de datos. Si bien en primera instancia el sistema depende de un operador humano para determinar si una palabra representa una opinión Positiva, Neutra o Negativa a partir de un conocimiento inicial se puede calcular automáticamente el valor de términos intermedios pero con la flexibilidad de permitir distintos algoritmos u operadores humanos para mejorar la precisión de los cálculos.

El modelado de la percepción permite determinar arbitrariamente que percepciones se ponderarán más en cuanto a un concepto ya que como se explica la cuantificación es en general universal en algunos contextos particulares algunos términos representan una opinión positiva y en otros casos negativas.

Las estrategias propuestas en este documento representan las mejores opciones para hacer *Opinion Mining* ya que son soluciones que pueden manejar un gran volumen de datos a partir de inferencias estadísticas, permiten una cuantificación uniforme, permite escalar de manera prácticamente ilimitada y al partir del modelado permite la mejora continua en sus puntos clave tales como la medición

de distancia entre conceptos, las definiciones de los valores Positivos, Negativos y Neutrales de los clasificadores, la clusterización y la búsqueda de patrones en los sondeos de opinión que es la clave del *Opinion Mining*.

VII. Referencias

[1] Susan T. Dumais (2005). "Latent Semantic Analysis". Annual Review of Information Science and Technology

[2] Ke Hu y Wing Shing Wong, "A Probabilistic Model for Intelligent Web Crawlers"

[3] Vladislav D. Veksler, Ryan Z. Govostes, Wayne D. Gray; "Defining the Dimensions of the Human Semantic Space"

[4] Rudi Cilibrasi y Paul Vitanyi, "The Google Similarity Distance", 2004

[5] The Google Similarity Distance, IEEE Trans. Knowledge and Data Engineering

[6] G. Jeh y J. Widom. SimRank: a measure of structural-context similarity.

[7] Gerlof Bouma, Normalized (Pointwise) Mutual Information in Collocation Extraction

[8] A. Evangelista and B. Kjos-Hanssen, Google Distance Between Words.

[9] Levenshtein VI (1966). "Binary codes capable of correcting deletions, insertions, and reversals"

[10] Navarro G (2001). "A guided tour to approximate string matching"

[11] Lovins, J.B. "Development of a Stemming Algorithm". Mechanical Translation and computation Linguistics.

[12] Andrews, K. "The Development of a Fast Conflation Algorithm for English". University of Cambrigde, 1971

[13] M.F.Porter, An algorithm for suffix stripping, 1980

[14] http://code.google.com/p/boilerpipe/

[15] http://jsoup.org/

[16] Bing Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)", 2009

[17] W. J. Frawley, Piatetsky G. Shapiro, C. J. Matheus, Ai Magazine, Vol. 13 (1992), pp. 57-70

[18] Freddy Y. Y. Choi (2000). "Advances in domain independent linear text segmentation"

[19] Christopher D. Manning, Hinrich Schütze: Foundations of Statistical Natural Language Processing

[20] Kirill A. Sorudeykin A Model of Spatial Thinking for Computational Intelligence

[21] Mihaiela Lupea, Doina Tatar, Zsuzsana Marian "Learning Taxonomy for Text Segmentation by Formal Concept Analysis"

[22] C. Schommer, "An Unified Definition of Data Mining", 2008

[23] http://incubator.apache.org/opennlp/

[24] Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine.

[25] Debajyoti Mukhopadhyay, Sajal Mukherjee, Soumya Ghosh, Saheli Kar, Young-Chon Kim "Architecture of A Scalable Dynamic Parallel WebCrawler with High Speed Downloadable Capability for a Web Search Engine"

[26] Dimitris Papamichail and Georgios Papamichail "Improved Algorithms for Approximate String Matching"

[27] Apache Software Foundation. http://www.apache.org

[28] Ludwig Bertalanffy "general system theory: foundations, development, applications (revised edition)", 1969