

**UNIVERSIDAD ANDRÉS BELLO**

**FACULTAD DE INGENIERÍA**

**INGENIERÍA CIVIL INFORMÁTICA**

**Trabajo 1 Etiquetado – Características Demográficas y Sicográficas en Usuarios de Plataformas cQA**

**GUSTAVO ANDRÉS VÉLIZ LÓPEZ**

**SANTIAGO-CHILE**

**MARZO, 2017**

**INTRODUCCIÓN**

Es necesario saber un poco sobre las comunidades de Pregunta-Respuesta para comprender mejor el desarrollo de este informe.

Las comunidades Pregunta-Respuesta (cQA) son sitios web donde los miembros de él pueden realizar preguntas para sean respondidas por el resto de la comunidad, así como también puede responder preguntas hechas por otros miembros. Un ejemplo de estas comunidades es Yahoo! Answers.

Dentro de estas *cQA* podemos identificar los siguientes elementos:

*Preguntas:* Generalmente presentan un título y un cuerpo, estas pueden pertenecer a cualquier categoría disponible en la comunidad.

*Respuestas:* Respuestas de usuarios a las preguntas disponibles en la comunidad, generalmente hay ciertas normas a seguir al momento de responder que dependen de cada comunidad.

*Categorías:* Las preguntas realizadas caen en categorías existentes en la comunidad, lo que permite un mejor orden facilitando la búsqueda de preguntas dentro de ella.

*Calificaciones:* Dentro de la comunidad es posible evaluar las respuestas hechas a alguna pregunta, lo cual permite discernir las respuestas útiles de las respuestas hechas en broma o equivocadas.

Gracias a estas comunidades, un usuario cualquiera puede encontrar respuestas ya existentes a preguntas, por muy específicas que sean, que pueda tener, lo cual atrae a personas provocando que la comunidad crezca.

Debido a la dimensión de la comunidad, al momento de realizar análisis de datos existan diversas formar para visualizar, agrupar y encontrar similitudes en los miembros de ella.

**DESCRIPCIÓN DEL PROBLEMA**

Para la Tarea 1 el alumno debe etiquetar manualmente un conjunto de perfiles de la cQA Yahoo! Answers. Para esto se les entregara en una página web las preguntas y respuestas que el usuario ha hecho públicamente, su nombre de usuario, una foto de perfil y una auto-descripción si están disponibles.

En base a esta información, el alumno deberá identificar y seleccionar desde una lista de variables tanto demográficas como sicográficas aquellas que sean acorde a como el usuario se define a sí mismo.

Luego de la realización del etiquetado, el alumno deberá obtener la probabilidad de cada variable, la entropía y la información mutua del mismo usando las formulas presentadas en la *Figura 1.*

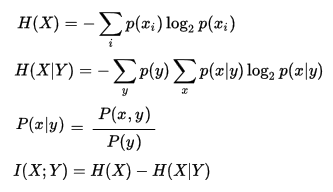


Figura . Ecuaciones de Entropía e Información Mutua

Siendo H(X) la Entropía de una variable, I(X;Y) la Información Mutua entre dos variables, y por último H(X|Y) y P(x|y) la entropía condicional y el porcentaje de una tabla con dos variables.

Para la realización de este informe el conjunto de datos consistía de 200 perfiles.

**DESARROLLO**

**Describe extensamente su experiencia de etiquetado: ¿Que observo? ¿Qué dificultades encontró? ¿Que observa de las preguntas que fueron en la clase “OTROS”/“NO IDENTITY”? Discuta esto en extenso y ejemplificando sus observaciones.**

Al momento de etiquetar se podían ver muchos perfiles similares en su estilo de vida, es decir, muy pocos perfiles eran de personas que no siguen una “vida tradicional”. Trabajamos exclusivamente con perfiles en inglés y español, siendo el inglés la lengua predominante, además, la gran mayoría de los perfiles no tenían información en su “descripción de perfil”, lo que provocaba que toda la información que se podía obtener de ellos, debía ser buscada en las preguntas y respuestas que fueron realizadas. En mi caso puedo decir que apareció al menos un perfil de cada edad. Esta se veía reflejada generalmente en su manera de escribir, el tipo de preguntas que se realizaban y en pocos casos, aparecía su edad en la descripción.

La etiqueta más difícil de obtener fue la de *Clase Social,* dado a que solo un pequeño número de personas comentaban sobre sus trabajos y finanzas abiertamente. En muchas ocasiones las preguntas y respuestas de las personas no otorgaban ninguna información útil para el etiquetado, como por ejemplo, personas que solo preguntaban sobre una aplicación, un juego o temas de ocio.

Respecto a aquellos perfiles que clasificaban en la clase “OTHERS” o “NO IDENTITY”, podemos decir que especifican no pertenecer a grupos disponibles en las etiquetas (los cuales clasifican como “OTHERS”), o que sus preguntas son centradas en un tema específico el cual revela poco o nada sobre ellos mismos (los cuales clasifican como “NO IDENTITY”). En el conjunto evaluado se encontró un gran número de perfiles así, siendo principalmente perfiles dedicados a resolver preguntas, es decir, solo tenían respuestas en su perfil público, o que preguntaban solo por temas de ocio, como encuestas de revistas juveniles, videojuegos o preguntas científicas.

**¿Qué características tienen los usuarios relacionados con cada valor de variable? ¿Cuál es la distribución en su conjunto de datos? ¿Cuál es la probabilidad de cada uno de los valores de variables?**

Dentro los 200 perfiles contenidos en el conjunto de datos evaluado, hubo 28 de ellos en los cuales no se pudo identificar ninguna característica etiquetable, es decir, que todas sus etiquetas quedaron como NO IDENTIFICABLE. En base a esto, se puede inferir que estos perfiles solo tienen interés en realizar preguntas que no involucren hablar de sí mismos, como puede ser el caso de pedir ayuda a la resolución de un problema o preguntar sobre algún tema de ocio o de interés común., así como también pueden responder preguntas similares.

Otra relación considerable era que la mayor parte de los usuarios entre 12 y 17 años de edad eran solteros, podía existir un par de excepciones a esto por variedad cultural o porque hayan mentido respecto a su edad y sus respuestas dieran a entender que tenía alguna clase de relación reconocida como estado civil pero no identificable en base a las mismas. Otro punto a destacar respecto a este grupo es que casi un cuarto de los usuarios de este grupo etario se consideraban a sí mismos como personas amigables. También destaca que al menos 6 perfiles dentro de este grupo se consideran personas creativas, característica que se puede asociar a la edad. Respecto a su Etnia, casi la mitad de este grupo no entregaba información que permitiera identificar su etnia, luego de esto, un poco menos de un cuarto del grupo era de etnia blanca y otro cuarto del mismo era latina o hispana, lo cual dejo a una minoría de 2 personas que pertenecían a la etnia negra y 2 más que pertenecían a la asiática.

En los usuarios entre 18 a 24 años se repite la característica de persona amigable en un tercio del grupo. La gran parte de los perfiles no mostró información de su estado civil, además dentro del grupo hubo un número similar de personas de raza blanca, negra y latinos.

A continuación se mostraran las distribuciones de las variables para cada etiqueta, eso sin considerar la auto-descripción.

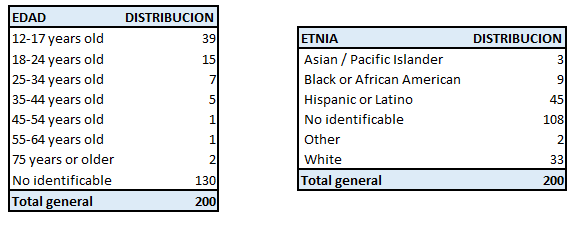


Figura . Distribución Edad y Etnia

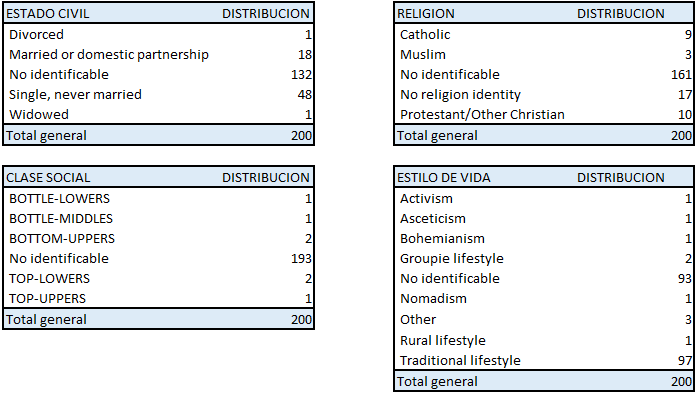


Figura . Distribución Estado Civil, Religión, Clase Social y Estilo de Vida

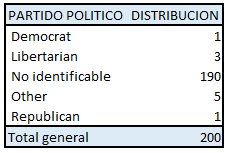


Figura . Distribución Partido Político

Por último, se graficaran las probabilidades de aparición que tienen esas variables en el conjunto.

Gráfico . Probabilidades de variables de Edad

Gráfico . Probabilidades de variables de Etnia

Gráfico . Probabilidades de variables de Estado Civil

Gráfico . Probabilidades de variables de Clase Social

Gráfico . Probabilidades de variables de Religión

Gráfico . Probabilidades de variables de Estilo de Vida

Gráfico . Probabilidades de variables Partido Político

**¿Hay perfiles comunes? ¿Logra identificar algún tipo de usuario dentro de la comunidad Yahoo! Answers? Elabore sobre este último punto.**

En el conjunto de perfiles existen varios elementos comunes entre ellos, pero el principal factor en común que posee casi la mitad de los usuarios es un *estilo de vida tradicional*, y esto aplica para personas de todas las edades, estatus sociales, religiones y tendencias políticas.

Por lo tanto, sin considerar usuarios que no pueden ser identificados en las etiquetas, se puede concluir que la tendencia de usuarios que conforman la comunidad evaluada, tienen un rango de edad que va desde los 12 años hasta los 24 años, que definen su estilo de vida como una vida tradicional y que tienen en común las características predominantes de ser amigables, abiertos de mente y sociales.

De los 200 usuarios evaluados existen 56 que en sus preguntas y respuestas no mostraron ningún dato relevante para la obtención de características demográficas, sin embargo, de estas 56 personas, 19 dieron en sus descripciones o en sus preguntas/respuestas datos útiles para identificar ciertas características sicográficas, más específicamente el cómo se describían a sí mismos.

**¿Hay relación entre ocurrencias de valores para las primeras cuatro variables (demográficas) y el segundo grupo (sicográfico)?**

Si existe cierta relación entre estos dos grupos, por ejemplo como se mencionó antes, hay una tendencia a que los usuarios entre 12 y 17 años, que por razones obvias son *solteros, nunca casados*, compuestos en su mayoría por personas de etnia blanca e hispánica y que no dan sean más *amigables* y *creativos* que el resto de los usuarios. Otro ejemplo es la tendencia que tienen los usuarios que indican que se casaron hace poco de ser personas *amorosas* o *afectivas*.

Pero como se mencionó antes, esto es una tendencia, dado a que hay excepciones dentro de estos grupos. Por ejemplo, si analizamos a las personas de religión católica veremos que solo 2 de ellos se consideran abiertos de mente y que podrían llegar a aceptar sin problemas otra religión.

Por lo tanto, la relación entre valores demográficos y sicográficos existe y nos permite agrupar grandes grupos de personas de modo que se pueda tener una idea de cómo será la actitud de una multitud que cumpla con las características demográficas, pero esto no es algo que se cumpla en el 100% de los casos.

**Calcule la entropía de sus conjuntos de datos para cada variable. ¿Cuál es la que tiene mayor certidumbre en la comunidad?**

A continuación se presenta un gráfico con las entropías de las variables con elección única.

Gráfico . Entropías de las Etiquetas

En base al gráfico, podemos ver que la mayoría de los valores son superiores a 1, esto se debe a que el conjunto evaluado considero 100 perfiles extra dando un total de 200 perfiles, en el caso que solo se hubiesen evaluado los 100 perfiles originales, los valores de la entropía oscilarían entre 0 y 1.

Cabe destacar que *a “menor entropía, mayor certidumbre”*, por lo cual se puede asegurar que la etiqueta que tiene mayor certidumbre en es la de Clase Social. En este conjunto de datos muy pocos perfiles daban información útil para determinar la clase social del usuario, lo que causo que hubiese un gran número de NO IDENTIFICABLE en esta etiqueta, llevando así al resultado mostrado en el gráfico

La siguiente etiqueta más certera es el Partido Político, el cual presento una situación similar a la Clase Social, con 190 perfiles que no demostraban ninguna tendencia política.

Por otro lado, la Etnia fue la etiqueta con entropía más alta, por lo que se puede concluir que el número de usuarios de diferentes etnias eran similares.

**Calcule la Información mutua para los valores de la variable edad y estilos de vida.**

Para realizar el cálculo de la información mutua entre estas dos variables, debemos usar la fórmula descrita en la *Figura 1*. Para esto necesitaremos obtener de la tabla los siguientes datos:

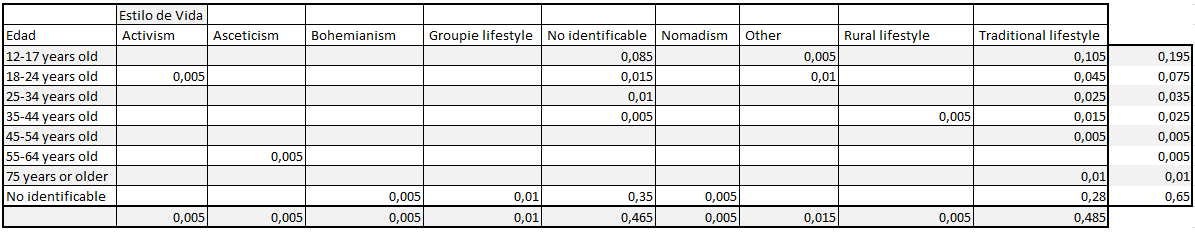


Figura . Tabla de Información Mutua

*Entropía de Edad: H(edad) = 1,589*

*Entropía Condicional entre Edad y Estilo de Vida: H(edad|estilo de vida) = 1,153*

Usando estos datos podemos obtener la Información Mutua de las variables edad y estilo de vida:

*I(edad|estilo de vida) = 0,436170697*

**CONCLUSIÓN**

En conclusión, luego de realizar el etiquetado y análisis de los datos se pudo determinar que:

Existe un gran número de perfiles que no proveen información de sí mismos (es decir que no usan la sección de auto descripción provista por la cQA), esto, si bien puede ser una buena medida de precaución por parte del usuario, dificulta el trabajo al realizar el etiquetado ya que provoca que la mayoría de los campos deban ser llenados con *No Identificable*.

Por otra parte, se pudo apreciar ciertas tendencias entre los usuarios evaluados, como que la mayor parte de la comunidad consiste en personas menores a los 25 años que son generalmente amigables y que se consideran personas creativas y/o abiertas de mente.

Esta información es útil para realizar publicidad o productos orientados a estos grupos, ya que pueden crear un perfil del cliente al que quieren llegar en base a las relaciones de las características demográficas con las sicográficas.

Por último, gracias al cálculo de la entropía pudimos identificar cuál de las etiquetas tuvo un mayor índice de certidumbre.