ICANN79 | Foro de la comunidad – Siguiente ronda del Programa de Nuevos gTLD: sesión de trabajo para la planificación de una revisión de similitudes entre cadenas de caracteres
Martes, 5 de marzo de 2024 – 3:00 a 4:00 SJU

PITINAN KOOARMORNPATANA: Hola, bienvenidos a la sesión de trabajo sobre la revisión de similitud de cadenas de caracteres, mi nombre es Pitinan y soy coordinadora de la participación de esta sesión, que será grabada y se rige por los estándares de comportamiento esperado de la ICANN. Durante la sesión las preguntas o comentarios serán leídas en voz alta, si se presentan en el pod las leeré en el tiempo asignado; si el tiempo lo permite, si desean hacer una pregunta o comentario verbalmente por favor levanten la mano, cuando diga su nombre tendrán permiso para abrir el micrófono. Hay interpretación en la sesión en inglés, francés y español, por favor digan su nombre para el registro y el idioma en el que hablarán, si no es el inglés, por favor hablen a un ritmo razonable y con esto le paso la palabra a Sarmad Hussain.

SARMAD HUSSAIN: Gracias, Pitinan. Buenos días, buenas tardes y buenas noches a todos, en esta sesión trataremos el trabajo de evaluación de las similitudes entre cadenas de caracteres, les voy a dar un contexto antes de comenzar. Este es un trabajo que forma parte esencialmente de la próxima ronda de solicitudes de nuevos gTLD como parte del Proceso de Desarrollo de Política, la evaluación inicial de cadenas de caracteres

claramente será sometida a una revisión de similitud de cadenas de caracteres por un panel.

El Proceso de Desarrollo de Política expeditivo con IDN, la fase 1, formuló recomendaciones que solicitan específicamente ese desarrollo y una guía de revisión de similitud de cadenas de caracteres para poder hacer un proceso más transparente para los solicitantes y la comunidad, entonces el trabajo que hemos estado conduciendo como parte de la implementación de estas guías de procedimientos posteriores, y también con la orientación que surge del EPDP fase 1, está comenzando a desarrollar estas guías.

Tratamos de resolver dos desafíos, uno que debemos determinar que exista un mecanismo claro o transparente para determinar qué es similitud de cadenas de caracteres para que todos entiendan y el segundo desafío que estamos intentando resolver aquí es que, con el advenimiento de los TLD con variantes, el ritmo al que hay que hacer la evaluación de las similitudes de las cadenas de caracteres ha incrementado significativamente, por ello quizás no sea factible hacer una revisión manual completa, entonces lo que estamos proponiendo es que haya un proceso de filtrado previo que se haga mecánicamente sobre la base de datos recabados entre las comunidades.

Eventualmente esto llevaría al panel de revisión de similitud de caracteres, tras un proceso de pre-selección, a que redundaría en un conjunto más pequeño de etiquetas que se puede ir a revisar

manualmente. Esta es la primera fase de estas guías, están publicadas para comentario público, está el enlace en la presentación a este comentario y presentan los factores que determinan la similitud que pueden utilizarse para determinar la transparencia, pero, además, un proceso de dos pasos, el que acabo de mencionar para abordar la escala que tiene este trabajo de manera más eficiente.

Hemos trabajado con el equipo que está desarrollando esta guías y ahora le paso la palabra a Michael Suignard, que es Fellow de Unicode y ha trabajado con nosotros en la elaboración de estas guías. Le pido por favor, Michael, que nos cuente un poquito sobre las guías, si ustedes tienen preguntas no duden en levantar la mano y podemos ir interrumpiendo a un ritmo razonable. Gracias.

MICHAEL SUIGNARD: Quiero decir que formé parte del panel de integración en representación de la ICANN, o sea que básicamente estoy trabajando en las LGR de la zona raíz, trabajé con el panel de generación de LGR, con ese grupo de expertos, así que, tengo cierta experiencia con las variantes y sé cómo manejar el contexto de la similitud de cadenas.

Las variantes soy muy importantes en este estudio, como veremos en la presentación, por un lado, asumimos que lo primero que se presenta cuando se solicita una etiqueta es que no se verán variantes porque habrán sido eliminadas como primera parte, creo que luego serían otorgadas a la misma entidad, esto sería un proceso de dos pasos. Hay

algunos conceptos que tendríamos que establecer, establecer la terminología, qué es confusibilidad, qué es similitud, qué son variantes.

Aquí hay algunos ejemplos que pueden ver, por ejemplo, lo que nosotros consideramos como iguales es algo que se visualiza como lo mismo, en este caso, la "C" de cirílico y estas variantes son eliminadas en la zona raíz como variantes, son visualmente similares en la siguiente fila, la "C" cedilla y el mismo símbolo en griego son ejemplo del Devanagari y el bengalí, son visualmente similares y no son variantes.

La última línea obviamente muestra cosas que no son visualmente distintas, así que, no tiene sentido analizarlo porque son distintos. Lo primero, la confundibilidad, es el grado en el cual dos etiquetas; no siendo diferentes, se pueden confundir, al determinar los confundibles podemos utilizar el término "confundible" tanto para adjetivo como sustantivo, entonces sucede cuando el grado de confundibilidad excede determinado umbral y este umbral es importante porque no es blanco o negro, no se decide el umbral según el propósito que, en este caso, es una LGR de zona raíz, es según como se aplica al contexto.

Tenemos cierta restricción, algunas de estas cadenas pueden ser sumamente cortas, la zona raíz es frecuente a tener dos letras, la cadena de dos letras o incluso una, es posible ver algunas que tienen un carácter. Por ejemplo, el chino puede ser la situación donde exista

un único caracter y a veces se parece a una sílaba similar, esto incluye los dominios ASCII, que es un marco muy diferente que hicimos para las LGR de la zona raíz, no consideramos los dominios ASCII o no consideramos variantes con ASCII porque no es posible básicamente.

Acá vemos la similitud, no se ve demasiado en la zona raíz, puede abarcar letras que cambien dependiendo del tipo de letra, si es la "I" cirílica, que no se parece a nada en el latino; lo mismo cuando se aplica en itálico, se parece a una "U". Lo mismo sucede con la "F" con el trazo o sin el trazo, fíjense que hay algunos límites de lo que es una variante y visualmente similar, a veces hay una zona gris entre lo que es visualmente similar y una variante, entonces el panel de generación ha considerado que algunos caracteres, que son visualmente similares, debieran ser considerados como variantes.

Las variantes, en ese sentido, son un poquito diferentes, es interesante establecer qué es una variante en comparación con un caracter visualmente similar, con una variante se estima que la etiqueta con variante es la misma y no está relacionado con la similitud visual o, desde el punto de vista visual, permite tener una única o múltiples letras y entre códigos de escritura se puede aplicar, no solo al aspecto visual, sino también al significado o a la pronunciación.

El caso más difícil de semántica sucede con el chino donde podemos tener variantes entre el chino simplificado y el tradicional, también en otros códigos de escritura donde aplicamos sonidos fonéticos para

aplicar las variantes, eso no lo hacemos en similitud, la similitud es estrictamente visual. Y, nuevamente, como hay equivalentes en variantes, existen las simétricas y las transitivas, por ejemplo, un trío, ABC en variante, si A es igual a B, B es igual a C y ABC en conjunto constituyen un término único similar a A y B, B y C, C y A, entonces la disposición de las variantes es un bloque.

Esto significa que las etiquetas con variantes no son permitidas si se usan los términos originales, se pueden, en algunos casos, buscar y hay delegaciones múltiples de la etiqueta original más las etiquetas con variantes. Esto sigue reglas estrictas en la LGR de la zona raíz para limitar el número de estas instancias y esto se aplica, en general, a chino o a árabe que tienen múltiples casos de estas situaciones asignables y por eso también usamos este término de etiqueta original o primaria, que es básicamente la que aplicamos.

En gran parte de estas guías no hay mucha diferencia entre la variante y la etiqueta original, hay que examinar después todos los estudios porque no hay diferencias significativas entre la etiqueta original y las variantes, por ejemplo, como ven en pantalla, la “f” en minúscula y la “f” con el trazo, y ahí hay variante. Cuando hagamos la revisión de similitud tendremos que trabajar en el reconocimiento dependiendo del contexto, es importante porque si se está evaluando, nuevamente, la similitud no se puede hacer en el vacío, hay que considerar el contexto.

Primero, tenemos que ver qué es un usuario, dependiendo a quién se le pregunte habrá distintas respuestas de lo que es un confundible, siempre asumimos que si estamos trabajando con un código de escritura tenemos que ser usuario del código de escritura, o al menos estar familiarizado, entonces tiene que haber algún tipo de subconjunto ASCII porque la zona raíz en su mayor parte no está disponible más que en ASCII, el 80% de la zona raíz utiliza cadenas de escritura ASCII y también está la cuestión de cómo reconocer las verdaderas etiquetas, no es necesario que estén presentes en el mismo tiempo, no tiene que estar en el mismo lugar, solo hay que estar familiarizado con uno y en otro contexto se parece con el que uno está familiarizado.

También sucede entre códigos de escritura o entre lenguas, no se espera que se puedan reconocer la similitud en códigos de escritura que uno no conoce, con los cuales uno no está familiarizado, no es un requisito, por ejemplo, alguien de Europa no se espera que reconozca las diferencias en texto chino al mismo nivel, no es caracter por caracter… [Sin interpretación]

Normalmente la similitud tiene que ver con un espacio de percepción donde no están en el mismo local, aunque parezca que está en el mismo local, por ejemplo, si uno tiene caracteres triples similares, el del medio puede parecerse a los dos que están en los extremos, pero los de los extremos tal vez no sean similares, entonces eso es lo que llamamos el espacio de percepción. De nuevo, las variantes no se

limitan a similitud visual, esto puede crear sus propios retos y vemos que las variantes no consideran la situación de mayúscula y por eso hay que considerar eso en cuando a la similitud, repito, eso puede crear ciertos retos.

En cuanto a las variantes pueden ser similar visualmente, aunque no fueron hechos para eso, sino para detectar si se ven iguales, hay un área gris, pero sabemos que esto no aplica a todas las situaciones similares. Ahí es cuando nosotros podemos ver que podemos crear algo similar a lo que llamamos un conjunto de variantes, básicamente esto es un conjunto que incluye a todas las variantes, puntos de códigos o cadenas de caracteres que se vean iguales, es establecer reglas para una percepción donde se vea un mapeo para una secuencia y determinar qué es lo que se puede hacer con respecto a eso y utilizar el sistema para monitorear esto.

Ahora, en cuanto a grupos confundibles, pueden ser grupos o etiquetas que pueden ser confundibles y así crear un espacio de percepción, se pueden crear muchas cosas confundibles hasta cierto grado, por eso tenemos que tener cuidado respecto a eso, hay un poco de experimentación que se hace en cuanto a eso para poder conseguir un mecanismo de preevaluación, pero podemos usar el sistema de variantes para que nos ayude en ese aspecto.

La interacción con la similitud de cadenas de caracteres con las variantes tiene que ver cuando uno compara un conjunto delegado,

hay que considerar las variantes de esos conjuntos delegados, lo mismo con las etiquetas de candidatos, entonces, como dijo Sarmad, básicamente es imposible resolver esto al hacer una revisión manual porque las variantes pueden ser infinitas en muchas situaciones, se pueden crear con etiquetas interesantes y no se puede resolver, de esa manera se hace hoy, por ejemplo, tenemos 1.600 TLD y eso es mucho, aún así no se puede hacer de una manera manual.

Hay una situación también donde se está mezclando lo confundible con las variantes en cuanto a etiquetas, la primera palabra es la palabra en francés para “avión”, lo mapeamos con ASCII con el acento en la “E”, en este caso, no es una variante, sino que es similar y después en la parte de abajo uno está mezclando una variante que se ve “AE” junto, pero puede existir una similitud común en cuanto a esto, por ejemplo, la ligadura de “OE”, esa puede ser también una variante similar.

Eso lo vemos, por ejemplo, en la palabra “coeur”, que es corazón en francés, donde es C-O-E-U-R, la “E” no forma parte de la ortografía en francés desde hace mucho tiempo, entonces es común que se transforme en las dos vocales, entonces ahí es donde se mezcla la variante y se puede confundir en este mismo nivel. Siguiente diapositiva, por favor.

Aquí también es muy importante que nosotros determinemos un nivel de similitudes, que eso nos puede ayudar en la preevaluación,

básicamente nosotros tenemos cinco niveles que se consideran idénticas o casi idénticas en el sistema de variantes, el número 2 son etiquetas altamente confundibles, se considera como una variante, por ejemplo, la “F” con un ganchito y después llegamos como a una zona gris, que son etiquetas similares que muchas veces no son variantes, en algunos casos sí, pero la mayoría de las veces son similares.

Después tenemos las etiquetas distantemente similares; donde se cree que es similar, pero la mayoría de las veces no, pero todavía le damos peso a eso. Luego existen las etiquetas distintas, en los ejemplos se ve que el primero es un ejemplo de lo mismo en latín y cirílico, donde utilizan las mismas letras y tienen la misma apariencia, pero es un significado completamente diferente, después en el latín se ve la pequeña diferencia de la letra “P” comparado al griego. Otro ejemplo es la P, A, Y entre el latín y el cirílico porque la “Y” griega realmente no es “Y” griega en cirílico. El otro ejemplo donde uno trata de utilizar ASCII, pero utiliza caracteres chinos, no se puede reconocer si es un nivel de dominio o no.

Una consideración interesante es que, a veces, hay letras sencillas como O, S, L, X, que es común para muchos códigos de escritura, las que se ven como un círculo. Si uno tiene un dominio que simplemente utiliza “O” eso se puede confundir con los códigos de escritura porque esa vocal existe en algunos códigos de escritura, por ejemplo, la palabra “coco”, se puede crear en múltiples códigos de escritura

donde no tiene un entendimiento común, pero visualmente se ve igual, por eso tenemos que estar al tanto de esto y es peor cuando tiene letras repetidas, como la “O” que lo tenemos en la raíz.

Hay uno que existe, que es triple “O” y esa triple “O” está bloqueando muchas variantes en esos códigos de escritura, entonces eso también se tiene que considerar, especialmente cuando el latín tiene muchas variantes cirílicas y está obstruyendo mucho de esto porque el ASCII está presente en eso. Este es un ejemplo muy rápido, no es necesario ver todo, a la izquierda vemos el latino “ENG” abajo del griego y a la derecha creo que es el ETA (êta = η), de alguna manera muestra lo que es una situación basada en el espacio de percepción, dos caracteres que se ven distintos de los costados porque desciende el brazo de manera distinta, mientras que el del medio tiene una pata recta, esto muestra lo que podría ser el espacio de percepción y es un ejemplo válido porque todos son considerados variantes, pero básicamente es una LGR de zona raíz.

Las mayúsculas y minúsculas. La IDN de 2008 elimina la combinación de mayúsculas y minúsculas de los navegadores, la combinación de mayúsculas genera confusión entre algunos códigos de escritura, en particular el griego y el latino porque en griego tenemos diferentes significados, esto es un ejemplo. La “H” en latino de repente pasa a la “N” en griego, luego pasa a la “H” en latino, después cambia a una “V”, de repente tenemos tres letras ASCII que son similares entre sí con mayúsculas, eso obviamente no se podría hacer, de ninguna manera

existe forma de tener una situación de bloqueo entre tres letras ASCII. Esto muestra el límite de lo que se puede hacer cuando se acepta mayúsculas como nomenclatura.

Entonces tenemos que conocerlo y también aceptar que tenemos límites para evitar la confundibilidad en mayúsculas porque tenemos que darles crédito a los falsos positivos. Luego las secuencias son confundibles, que todavía no está tan explorado, aquí vemos el conocido ejemplo de RN en comparación con el “M”, que se ha usado en el pasado para phishing, nadie sabe lo que es un MySpace, pero en mi época se usaba MySpace para hacer este tipo de ataque.

Luego vemos secuencias relacionadas, como mencioné antes, las ligaduras y a veces tenemos características repetidas en el código de escritura de Myanmar, hay características que no se ven porque están ocultas en el medio, aquí tenemos un ejemplo de una secuencia que usted necesita para evitar una confusión. En Myanmar se pueden poner dos letras que parecen 60, en lugar de que sea al revés, que es el orden lógico. Este es el contexto a tener en cuenta al evaluar la similitud, hay que ver el contexto de código de escritura, entonces hay consideración que se específica de ciertos códigos de escritura. Siguiente.

Estamos asumiendo siempre que en una situación de confundibilidad el usuario objetivo conoce el código de escritura y esto es muy importante, así que, siempre lo vamos a asumir, no es que estamos

intentando crear conocimiento o consciencia global de la confundibilidad, tenemos, no obstante, que saber que la lectura por parte de un nativo también tiene un peligro porque el cerebro se utiliza para generar un corto circuito, a veces las cosas son iguales, como vemos después.

También se basa en este concepto que depende del ojo de quién lo mire, la buena noticia es que se ha estudiado mucho, algunos de los paneles de generación de códigos de escritura han evaluado lo que puede ser considerado visualmente similar. Hay algunos códigos de escritura que están relacionados por su historia, sus características, como el latín, el griego, el cirílico y el armenio; que es sorprendente, pero estaban en la misma escena original. Lo mismo con chino, japonés y coreano, por coreano me refiero a Hanja; que es la versión más china del coreano, no Hangul, que es muy diferente.

Luego los códigos de escritura Neo Brahmi, los que se usan en India Occidental, que también tienen muchas estructuras comunes, no todos, pero luego veremos algunos ejemplos. Y algo más es que, los tipos de escritura de interfaz de UI intentan armonizar los glifos entre los códigos de escritura que de otra manera no estarían o no podrían ser armonizados. Hay algunos códigos de escritura correlacionados como Kannada y Telugu, que están muy coordinados, hay 34 formas comunes entre ambos; que es bastante alto, entre latino y cirílico que tienen 28, que es sumamente alto, sin contar acentos, y Devanagari y Gurmuki también.

Hay otro caso, como el tailandés y el Lao, que están muy relacionados desde su origen, pero, en realidad, su forma es muy diferente, así que, no sé si es tan válido. Siguiente.

Y luego tenemos algunos códigos de escritura con débil correlación, pero que, no obstante, requieren nuestra atención, vemos cierta confusión entre el Han, el Katakana y el Hiragana, pero estas confusiones ya han sido capturadas por el sistema en la zona raíz, también hay similitudes visuales que podemos ir viendo más si merecen un estudio adicional. En el caso del Han algo interesante, como no hay tantos problemas entre códigos de escritura como sucede con el árabe, el etíope, el Gujarati, el Khmer… No presentan tantos problemas entre cruces de códigos de escritura, hay algunos ejemplos en dos cadenas de caracteres, Kannada y Telugu, donde muchos pueden no notar las diferencias.

La situación con el armenio, este es un ejemplo entre la escritura tradicional del armenio, que no se ve para nada como latino, pero cuando tomamos la versión más moderna se parece más al latino, no en lo que hace a palabras… No sé si hay alguien armenio aquí…En algunos casos parece “U”, pero pueden crearse en armenio otros términos en latino y en griego, dependiendo del código de escritura que se utilice, puede generar más o menos similitud con el latino. El serif tradicional se ve menos, mientras que en el caso del serif moderno se ve más.

Y es interesante en el caso de la "π", a veces se puede encontrar confusiones según el formato, este es un ejemplo de donde obtenemos dos letras o una letra, donde se puede incluso generar confusión. Ahora entramos a la parte del proceso en nuestra discusión, no podemos aplicar la fuerza bruta porque existen demasiadas combinaciones casi infinitas, entonces la única forma de manejar esto es tener un proceso de dos pasos, el primero es mecanizar, en el cual utilizamos algo básico del sistema de variantes y luego una fase de post-selección donde un grupo de expertos en un panel evalúan el subconjunto de las etiquetas en la zona raíz para determinar si tienen o si son variantes.

A veces no es necesario, pero si se necesita hay que determinar que el alcance del trabajo sea manejable y el tamaño. Está bien, sigamos. El tamaño del registro objetivo también importa, yo tengo números que quizás sean distintos de los reales porque utilice los de etiquetas IDN y estos son otros, son aproximadamente 1.590 entradas, la mayoría, 1422, son entradas ASCII y 168 entradas de IDN. Esto muestra las variaciones de ASCII e IDN.

En el caso de las ASCII tienen muchas variantes que no está todo claro, pero existen, AAA, ARP, AC, ACO… Serían niveles cirílicos totalmente válidos, pero también otros códigos de escritura, encontré una entrada OPR en cirílico que sigue una secuencia latina perfecta. Si vamos por códigos de escritura la mayoría son latinos, otros dos más que los

ASCII, CJK, 66 en su mayor parte chinos, algunos japoneses y 36 árabes, y luego va bajando, algunos poquitos cirílicos, Canaán, japonés, Devanagari, Hangul, etc., porque creo que India creó una entrada para el nombre del país en cada código de escritura oficial, por eso básicamente aquí está representado el Devanagari, etc.

Como ven, el latín y el ASCII es el punto de análisis porque es la vasta mayoría, la similitud de códigos de escritura va a estar sobre todo centrada aquí, el tamaño que tiene la zona raíz entonces sugiere que debiéramos hacer un trabajo adicional. El plan es usar el sistema de variantes para hacer pre-selección en la zona raíz, como lo hacemos, lo usamos a través de lo que se llama “una etiqueta índice” o a veces “una etiqueta variante”, que en la tecnología informal se llama “esqueleto”, es decir, asociamos cada etiqueta asociada a la variante, se le asigna un valor único para no tener que comparar todo con todos, se asigna un valor y si hay dos valores se resuelve porque permite este mecanismo no tener que hacer una comparación interminable, se pueden comparar más según el universo.

El sistema de variantes entonces es muy bueno por esto, ¿sí?

BILL JOURIS: Disculpas, cuando leí el documento en preparación para esta sesión no vi nada sobre cómo se calcula este índice y si bien tengo mucha imaginación, no puedo imaginar cómo es posible calcularlo y obtener algo que sea significativo.

MICHAEL SUIGNARD: Siempre utilizamos el valor más bajo de cualquier variante, de paso digo que el índice no tiene que ser una etiqueta válida, no lo forzamos, pero básicamente es el punto de código más bajo posible en la etiqueta posible, entonces se toma cualquier conjunto de variantes y básicamente se usa el punto más bajo.

BILL JOURIS: Gracias.

MICHAEL SUIGNARD: Es bastante arbitrario, se puede usar el más alto, pero bueno, suele ser más fácil. La mayoría de las veces es una etiqueta válida, pero no siempre y algunos índices pueden no ser válidos porque aún no es un requisito que sea válido. Entonces hacemos lo mismo en la pre-selección, básicamente creamos clusters y luego los tratamos como un conjunto de variantes, y con variantes usamos lo que se llama "mapeo" entre el punto de código de variante o una secuencia, que es un mecanismo que permite generar una disposición en la zona raíz, aquí lo podemos usar como un mecanismo para hacer triage de los valores de las etiquetas para ver cuán cerca están en el espacio de percepción.

La idea es usar un mix entre el conjunto de variantes mapeando para determinar cuáles son disimilares, débiles, similares, etc., o sea se

pondera del mismo modo que se hacen las mediciones, etc. Obviamente lo malo aquí es que se pueden incluir algunas variantes que son confundibles… Permítanme, rápidamente, seguir con la siguiente diapositiva.

Entonces tenemos aquí datos que surgieron de los estudios de la zona raíz, grupos como, por ejemplo, el Latín GP dio información, pero no mucho en cuanto a la secuencia, es un dominio que no se exploró mucho. Y vemos los resultados de la revisión de similitudes entre cadenas de caracteres, que no fue elegible debido a la variante en cuanto a qué pasó, bueno, porque visiblemente fue algo bueno, hicimos una preevaluación y no vimos ningún problema, así que, estaba bien.

Con respecto a la contención es cuando tiene que ver con etiquetas que se solicitaron, si hay dos sería confundible entre sí mismos, en cuanto a si falla es cuando vemos una confusión con una etiqueta existente. Siguiente diapositiva.

Bueno, como no tenemos tiempo, más bien pasemos a la otra diapositiva. Ahora en cuanto a evaluación heurísticas usamos una puntuación de 1 a 5, básicamente utilizamos el mismo sistema, como la zona raíz, para visiblemente hacer la comparación, la parte de acciones para determinar qué tan precisa son en el espacio de percepción. Siguiente diapositiva… Siguiente diapositiva más bien.

Para que eso funcione tenemos que crear LGR, o sea tomar los 26 LGR que ya existen en la zona raíz y usar una situación diferente de mapeo o metodología, pero lo mismo aplicaría al aplicar las variantes y crear un nuevo conjunto de LGR que sean internos al mecanismo, y básicamente eso va a crear un sistema de grupos. Y, como dije anteriormente, usamos un índice de grupo, colectamos una serie de etiquetas y hacemos un análisis según ese grupo. Bueno, eso es mucho texto, no voy a resumirlo, entonces eso suele utilizar otro sistema y puede ser útil en algún contexto, en vez de usar un sistema de variantes más bien se utilizarían 18 pasos basado en Levenshtein, es otra forma de evaluar la similitud al usar este proceso de Levenshtein, uno puede añadir una letra o doble letra. Es un sistema que tal vez utilicemos para hacer esta preevaluación de similitudes, pero no tengo tiempo para explicarlo en este momento.

Entonces vemos que la post evaluación es una evaluación manual, la idea aquí es que van a utilizar mucha información creada por la preevaluación para determinar, primero, si el nivel que se está evaluando es pequeño y después evaluar la preevaluación. Después, para su propio nivel de revisión, de pronto inyectar su propia etiqueta y que no sea confundible para la preevaluación, pero pensamos que debe formar parte del grupo.

Esto le da cierta flexibilidad de añadir etiquetas en esa revisión de cadenas de caracteres, sería de miles a miles o en menos de eso, entonces es bastante fácil y se pueden ver estos principios de

preevaluación para hacer su propio estudio. El panel no va a utilizar grupos de enfoques, pero puede consultar a expertos de códigos de escritura. Entonces esperamos tener tres revisores en el panel de post evaluación, tienen que tener calificaciones en tener familiaridad con el código de escritura y seguridad del DNS, y tener la habilidad de anular cualquier puntuación de los resultados, si es que se puede rechazar si es confundible o aceptar y después tienen que documentar la lección, o sea, ¿por qué tomaron la decisión y en base a qué?

La siguiente diapositiva… Creo que ya terminé, pero bueno, los siguientes pasos a tomar son básicamente afirmar lo que estamos haciendo y después recibir comentarios públicos sobre el documento, tal vez hacer un ensayo del conjunto existente o tomar los 1.500 conjuntos que tenemos en este momento y aplicar esa revisión de similitudes entre cadenas de caracteres para ver si hay cuestiones similares o no, y después ver si hay un progreso. En la siguiente diapositiva creo que simplemente muestra los enlaces al cual uno ir.

SARMAD HUSSAIN: Bueno, gracias. Ahora es el momento para preguntas y respuestas.

PITINAN KOOARMORNPATANA: Bueno, para que sepan, si quieren alzar la mano virtualmente lo pueden hacer. Pido, por favor, que siga con su pregunta.

BILL JOURIS: A mí me gustan los cinco niveles de similitudes que tienen más o menos en la diapositiva número 40, pero no estoy seguro en cuanto a cómo se le asignan esos niveles, cuando el panel de generación latín estaba tratando de decidir si algo era una variante o no, nos tomó como cinco años para repasar todo eso o mirar todo eso y esto me parece que va a ser un esfuerzo similar y no estoy seguro cómo se va a lograr esto antes de que comience la siguiente ronda.

MICHAEL SUIGNARD: Es bueno que sepa que fueron cinco años, pero la intención es utilizar el trabajo que ya se hizo de la LGR, no queremos reinventar la rueda, ya se ha hecho mucho trabajo con respecto a esto y pensamos utilizar el trabajo que se ha hecho para crear o para evaluar eso. Hay decisiones arbitrarias, tal vez se tenga que usar otro código, pero las buenas noticias, tomando en cuenta que he estado participando bastante en esto, es que estamos al tanto de esto.

Hay ciertos códigos de escritura en los cuales tenemos más conocimiento, aún creo que hacemos bien en ver las similitudes de cadenas de caracteres y aunque exista disparidad pensamos en preguntarles a los expertos en cuanto a esto, por ejemplo, en la India o, por ejemplo, en cuanto al árabe de pronto haya un grupo que nos pueda ayudar con respecto a eso porque no pretendo tener conocimiento con respecto al aspecto árabe, pensamos pedirles a otros grupos que nos ayuden con eso. Muy bien, Alan.

ALAN GREENBERG: Gracias. Tengo dos preguntas, yo me confundí con la sección 9.1 y 9.2 que habla del tamaño de los registros porque esa es una cuestión de segundo nivel, entonces si pudiera contestar eso porque estoy un poco confundido. Bueno, haré las dos preguntas. Y la segunda pregunta es que, me sorprendió un poco que, aunque estamos fijando en lo que es similitud visual, ¿han ocurrido avances en el reconocimiento de patrones en los últimos 10 años? Y yo esperaba ver que se mencionara eso, aunque se hubiera descartado.

SARMAD HUSSAIN: Para responderle a su segunda pregunta hay una recomendación de no utilizar herramientas automatizadas para esto y obviamente hay que pasar por un proceso manual de revisión. Gracias.

ALAN GREENBERG: Bueno, yo entiendo por qué no queríamos utilizar la última, pero el mundo ha cambiado mucho desde ese entonces porque ya estamos en una época que su celular reconoce el rostro de uno, que no se podía hacer en el 2012.

MICHAEL SUIGNARD: Bueno, si mencionamos los estudios que se han hecho del reconocimiento de palabras y otros asuntos, entonces hay cierta discusión en las pautas con respecto a eso. Ahora en cuanto al tamaño del registro no sé si entendí bien su pregunta, pero solo nos fijamos en

la zona raíz no segundo nivel, esto es una lista de zona raíz, así que, yo no sé a qué se refiere cuando dice “segundo nivel”.

ALAN GREENBERG: Bueno, sección 9.1 y 9.2 habla acerca de la relevancia con un registro grande y es complejo, por eso no sé por qué se mencionó.

MICHAEL SUIGNARD: Es porque traté de utilizar… Se puede escalar en un registro más grande, pero por el momento no es el ámbito para este tipo de registro, tienes razón.

PITINAN KOOARMORNPATANA: Gracias. Nabil.

NABIL BENAMAR: Solo una pregunta con respecto a código de escritura árabe y como miembro del grupo de trabajo con respecto a IDN del código de escritura árabe, creo que hemos hecho trabajos para lo que se ha implementado en la zona raíz para el dominio de nivel superior y también para el segundo nivel para diferentes idiomas, entonces estoy curioso, ¿qué es lo que falta en este trabajo para que nosotros podamos revisar o revivir al grupo de trabajo?

MICHAEL SUIGNARD: Bueno, creo que este ámbito es diferente al ámbito anterior, estábamos mirando las variantes, solo las cinco categorías, pero ahora estamos viendo uno de los primeros tres, entonces como el ámbito cambia tal vez después del análisis el panel de generación o la comunidad de código de escritura tal vez piensen que no hay que añadir nada más o identificar para revisión de similitud, pero todavía se requiere que se revise una vez más para asegurarnos de este tema y si hay algún punto adicional que haya que identificar, que se pueda incluir esto también.

PITINAN KOOARMORNPATANA: Muy bien, gracias. Siga, Bill.

BILL JOURIS: Primero, quiero decir que estoy feliz con el documento que ustedes han hecho, yo tuve un sinnúmero de cuestiones en cuanto a nuestra posición anteriormente y los documentos han solucionado todas mis cuestiones, casi todas mis cuestiones, y estoy feliz al ver esto, sin embargo, sí creo que faltó algo, el subrayar. Cualquier browser principal, Chrome, Firefox, estos cuando uno automáticamente tiene un nombre de dominio puede cambiar el color, que no es problema, y lo puede subrayar, los procesadores principales como Microsoft Word, Acrobat, hacen lo mismo.

Entonces hay símbolos que tal vez tengan un problema con eso, con el subrayar, uno de los miembros de código de escritura del latín es la

"O" subrayada, si uno subraya un nombre de dominio que ya tiene una "O" subrayada, entonces pueden existir espacios en blanco, pero como ese caracter solo ocurre en lenguajes no tan comunes un usuario normal no va a saber buscar de que existe ese tipo de interrupción debido a la situación de subrayar la palabra.

MICHAEL SUIGNARD: Sí, estamos familiarizados con eso en la zona raíz, pero su punto tiene que ver con el contexto mayormente latino o cirílico, yo no creo que sea un problema en general, pero yo creo que es un punto válido y creo que deberíamos añadir el concepto de qué ocurre cuando se subraya la palabra, pero no nos hemos finado mucho, pero si la comunidad cree que hay que añadir ese aspecto, entonces sí está bien, no tengo ninguna objeción a eso.

SARMAD HUSSAIN: Gracias, Bill, creo que es una buena sugerencia. Y, como dijo Michael, buscamos aportes de la comunidad, entonces si creen que hay ciertas cosas que hay que añadir, por favor, pongan un comentario público y definitivamente lo consideraremos.

BILL JOURIS: Y yo entiendo que todavía está en progreso ese tipo de trabajo.

MICHAEL SUIGNARD: Hablamos de lo plural contra lo singular, aunque parece que es una vista céntrica y a veces eso no es cierto en inglés porque en las palabras francesas se utiliza el plural a veces, por eso esto es un concepto peligroso. Entonces esto es algo que tal vez consideremos, pero en este momento no se ha considerado y esto ha sido por su diseño.

PITINAN KOOARMORNPATANA: La última pregunta es de Ben.

BEN MCELWAINE: Soy Ben McElwaine de Google Registry, o registro de Google. Parece que esto tiene que ver totalmente con la similitud con códigos de escritura visibles, me pregunto si, ¿alguien ha estado trabajando en esto? Porque parece que existe otras confusiones potenciales que serían peores porque se basa solamente en caracteres latín, por ejemplo, si existe la palabra “coin” o “coins” hay muchos idiomas que no tienen el concepto de algo en plural, entonces si no hablan en inglés como el primer idioma y escuchan eso, o algo en plural, de pronto no se dan cuenta que es plural, entonces tal vez vaya a la versión singular y eso puede crear un problema.

Y también las variantes en la ortografía, cómo se escribe color en inglés americano comparado a cómo se escribe color en el Reino Unido, uno tiene la “U” y la otra palabra no la tiene, entonces, en este caso, tal vez no tiene que ver con esto porque estamos hablando de similitudes

visuales, pero ¿hay alguien trabajando en esto? ¿Están permitiendo que esto se delegue o lo están bloqueando?

SARMAD HUSSAIN: Gracias, Ben, por su pregunta. Hay una recomendación que sugiere evaluar lo que es singular y plural, aparentemente lo está considerando la comunidad de la Junta Directiva y, además, no hay otras recomendaciones para agregar ningún otro alcance de similitud, más allá de la parte visual de la similitud, se pueden tener similitudes fonéticas, gramaticales u otras. Además de números y poder tener variación de género o entre mayúsculas y minúsculas, todas estas variaciones lingüísticas y morfología por inflexión, pero este trabajo está siendo guiado estrictamente por el alcance de las recomendaciones de políticas que provienen… O sea, esta sección de lo que hay que hacer está estipulado por la política.

MICHAEL SUIGNARD: Usted mencionó un ejemplo de visibilidad, pero hay todo un universo, por ejemplo, dentro del francés y del inglés hay muchos términos que son distintos simplemente por pequeños cambios, por duplicar la consonante y uno se confunde todo el tiempo, hay muchos ejemplos. Usted mencionó la “O” y la “OU” en color, pero hay millones de ejemplos con distintas dimensiones, entonces tenemos que elegir la dimensión que seleccionamos, no podemos ser perfectos obviamente y al final siempre habrá un panel que tendrá una parte manual del proceso.

Si hay confusión en color, por ejemplo, en el proceso se puede resolver manualmente, la idea es simplificar con la pre selección, disminuir el tamaño del estudio para poder hacerlo manual porque no se puede hacer un estudio manual de todo.

SARMAD HUSSAIN: Bien, vamos a cerrar la sesión porque nos hemos excedido del tiempo, gracias por asistir, preguntar y participar. Gracias, Michael, por presentar y si tienen alguna pregunta adicional bueno, no duden en contactarnos en la dirección que conocen, lean el estudio y aquí estamos para el comentario público y vuestros aportes. Muchas gracias, con esto concluyo la sesión.

[FIN DE LA TRANSCRIPCIÓN]