UNIVERSIDAD TECNOLÓGICA NACIONAL

Facultad Regional Rosario

Maestría en Ingeniería en Sistemas de Información

Tesis de Maestría - Comentarios devolución

"Desarrollo de una medida de similaridad para Sistemas de Recomendación en sitios de Community Question Answering. Análisis desde un enfoque Big Data y usando un método de ensamble de clustering"

Ing. Federico Tesone

Director: Dr. Guillermo Leale

Co-director: Dra. Soledad Ayala

Rosario, Santa Fe, Argentina. Octubre de 2021

Contenido

En el dictamen de Maria Soledad Pera PhD. se destacan algunos puntos relativos ciertos aspectos del trabajo de tesis "Desarrollo de una medida de similaridad para Sistemas de Recomendación en sitios de Community Question Answering. Análisis desde un enfoque Big Data y usando un método de ensamble de clustering", con el fin de posicionar de una mejor manera las contribuciones realizadas con respecto al estado del arte. A continuación se detallan las propuestas observadas y se adjuntan, para cada una de ellas, los posibles cambios y/o comentarios que el tesista considera pertinentes para materializar dichas oportunidades de mejora.

Comentario 1

En el Capítulo 3 se hace mención al trabajo de Goldberg et al. (1992) como el primer sistema de recomendación, y se usa la citación al trabajo de Schafer el al. (2007) para definir sistemas de recomendación basados en collaborative filtering. En la comunidad, no son esas las citas más conocidas. Sugiero que el candidato considere las citas en este link, en lo que se refiere recommender systems y collaborative filtering: `https://paperpile.com/shared/OUBDis`.

Respuesta

Luego del párrafo que inicia el marco histórico de los RSs, que menciona y cita a (Goldberg et al., 1992) y del párrafo que menciona el problema de la sobrecarga de información, podria agregarse lo siguiente:

> "En la década de los 90s, fue de gran importancia la contribución de (Shardanand y Maes, 1995) como un enfoque para la resolución de la sobrecarga de información. En particular, se implementa un sistema llamado Ringo, que genera recomendaciones personalizadas para álbumes y artistas musicales. Dicha contribución, puede extrapolarse a recomendaciones de cualquier tipo de base de datos basados en similaridades entre usuarios.". Por otro lado, este trabajo también podría haber sido mencionado en la sección "3.2.3.5. Basados en comunidades", ya que la implementación descrita explota similaridades entre

> los gustos de los distintos usuarios a recomendar, y supone que los gustos, no están aleatoriamente distribuidos, sino que se encuentran patrones tanto entre persona como entre grupos sociales de usuarios. Se utilizan los usuarios con perfiles más similares para dirigir recomendaciones personalizadas.

Nota: en el mismo párrafo, se menciona lo siguiente:

> "Aunque las raíces de los RS se remontan a trabajos en ciencia cognitiva (Rich, 1979), teoría de aproximación... "

El cual es uno de los trabajos sugeridos por la Jurado. Una buena adición al trabajo podría ser dar detalles acerca del contenido de dicho trabajo, ya que es muy bien reconocido por la comunidad y sienta las bases para reconocer y diferenciar distintos tipos de usuarios que utilizar un sistema de información, creando estereotipos para representar con exactitud distintas características de usuarios y, a partir de ellos, recomendar novelas. Para la definición de los estereotipos, se captura información del usuario mediante preguntas personales. Este sistema, llamado Grundy, es pionero en RS basados en conocimiento.

El artículo (Sarwar et al., 2002) aporta un gran valor a este trabajo de tesis. Provee una gran definición para RS de filtrado colaborativo, la cual puede ser incluida en la seccion "3.2.3.2 Filtrado Colaborativo" como:

> "El objetivo de los algoritmos de recomendación basados en filtrado colaborativo es sugerir nuevos productos o predecir la utilidad de cierto producto para un cliente en particular, basándose en las elecciones anteriores del cliente, u opinión de otros clientes con gustos similares (Sarwar et al., 2002)".

En reemplazo de la cita de (Schafer et al., 2007).

Nota: Por otro lado, el concepto de Valor Singular de Descomposición (SVD), puede ser agregado al trabajo en sección "3.3.1. Contexto Histórico" de Big Data como un método de escalabilidad de procesamiento de datos mediante reducción dimensional.

El trabajo (Sarwar et al., 2001) que hace foco en filtrado colaborativo basado en ítems, es de gran valor para este trabajo, ya que se ubica directamente en el área temática, ya que utilizan una matriz de similaridad para explorar relaciones entre ítems (preguntas en sitios de CQA) y calcular recomendaciones a usuarios. Es un artículo que podría ser aplicado en la fundamentación de este trabajo de

tesis. Este artículo, por otro lado, aporta la distinción de que las opiniones de los usuarios pueden ser obtenidas *explícitamente* de los usuarios o utilizando algunas medidas *implícitas*, lo cual puede ser agregado a "3.2.3.2 Filtrado Colaborativo".

Con respecto a los trabajos (Hill et al., 1995) y (Resnick et al., 1994) tambien podrian haber sido agregado para ampliar el marco histórico de los Sistemas de Recomendación, y definitivamente si se hubiese ampliado el mismo para RS de filtrado colaborativo.

Comentario 2

FastText y Word2Vec son dos representaciones de embeddings. Mientras que la forma de producción de embeddings (proceso) difiere, el objetivo (representación) es común. Me pregunto entonces por qué presentarlos como subsecciones separadas en el Capítulo referente al marco teórico para este trabajo?

Respuesta

Hubiese sido una presentación completamente válida si los dos algoritmos se encontraban en la misma sección, indicando sus diferencias. Se optó por presentarlos de manera separada por los siguientes motivos:

- Se encuentran separados en el trabajo tomado como estado de arte. En el trabajo (Gonzalez et al., 2017), los autores describen y presentan los resultados de FastText y Word2Vec de forma separada. Esto conlleva a mantener la misma separación con un agregado de profundidad en cada uno de ellos.

- Si bien ambos algoritmos tienen las mismas características (modelos basados en redes neuronales que producen como salida un conjunto de palabras con representación vectorial), también tienen sus diferencias. La diferencia más importante (además de la implementación) es que FastText introduce el modelo sub-palabra en la cual cada palabra es representada como un conjunto de n-gramas, produciendo vectores con información sub-palabra, que han demostrado lograr mayor precisión en algunos casos.

- La sección "3.4.5. FastText" no explica nuevamente implementación del modelo Skip-gram (como sí lo hace la sección "3.4.4. Word2Vec"), sino que parte del mismo y agrega sus diferenciadores.

Comentario 3

En la definición de clustering, por ejemplo, el candidato utiliza (Peña, 2013) como referencia. Lo mismo ocurre en varias oportunidades en lo que se refiere a definir conceptos fundacionales, como Big-O. Es preferible en trabajos de esta envergadura que se utilicen las citaciones originales (o por lo menos a trabajos más reconocidos) para referirse a conceptos ya establecidos en la comunidad científica. Por ejemplo, un buen punto de partida es el libro Machine Learning, escrito por Tom Mitchel, o trabajos como "Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2), 103-134."

Respuesta

Con respecto a la definición de Clustering, adhiero que necesita citas más influyentes dentro de la comunidad. Se propone reemplazar la cita de (Peña, 2013), con lo siguiente:

> El Clustering o análisis cluster tiene como objetivo descubrir los grupos naturales en un conjunto de patrones, puntos u objetos. Los autores en (Jain, 2010) definen análisis cluster como "una técnica de clasificación estadística para descubrir si los individuos de una población caen en diferentes grupos haciendo comparaciones cuantitativas de características múltiples"[1] . Por otro lado, una definición de clustering operativa, es: dada una representación de n objetos, encontrar K grupos basados en una medida de similaridad tal que las similaridades entre objetos del mismo grupo son altas mientras que las similaridades entre objetos de diferentes grupos son bajas. Estos métodos se conocen también con el nombre de métodos de clasificación automática o no supervisada. Básicamente, los sistemas de clasificación son supervisados o no supervisados, dependiendo si asignan nuevos objetos de datos a uno o a un número finito de clases supervisadas discretas o categorías no supervisadas, respectivamente (Xu y Wunsch, 2008). Muchos enfoques han sido propuestos para manejar el problema de aprendizaje en presencia de variables no observadas. El algoritmo EM[2] es la base para muchos algoritmos de clustering no supervisados (Mitchel, 1997), entrenando un clasificador que utiliza documentos etiquetados, y probabilísticamente etiqueta los documentos no etiquetados. Luego, entrena el clasificador utilizando las etiquetas para todos los documentos, itera y converge (Nigam et al.,

[1] Definición obtenida por Webster [Merriam-Webster Online Dictionary 2008]. `https://www.merriam-webster.com/dictionary/cluster%20analysis`. Último acceso: Octubre 2021.

[2] Por sus siglas en inglés de Expectation-Maximization.

2000). En el momento de lidiar con documentos no etiquetados, se puede pensar a EM como un algoritmo de clustering no supervisado.

Por otro lado, se define *Big-O* desde el lado de las ciencias de la computación. Para lo cual, se utiliza el libro (Cormen et al., 2009), bandera para la definición y desarrollo de algoritmos de computación. Es posible agregar una definición matemática, la cual es fundacional (Paul Bachmann, 1894), como:

> Siendo $g, f : \mathbb{N} \to \mathbb{R}$ dos funciones de números naturales a reales. Decimos que f tiene orden de g, y escribimos $f = O(g)$. Si y sólo si hay una constante positiva c tal que por todos los suficientemente grandes valores $n \in N$ para $n \geq N(c)$, los valores absolutos de las dos funciones satisfacen la siguiente relación $|f(n)| \leq c\,|g(n)|$ (Erk y Priese, 2008).

Comentario 4

El alumno claramente detalla los inconvenientes que emergen al tener que lidiar con Big Data, me pregunto si se consideró un paso de candidate selection previo al análisis y recomendación, de manera de aminorar las comparaciones necesarias a medida que más preguntas se suman al sistema de CQA. En otras palabras, se asume en este trabajo que todo el conjunto de datos de entrada (pares de preguntas) debe ser analizado para generar clústeres y matrices de coasociación. Me pregunto si se consideró -en mira a la recomendación- que podía ser posible seleccionar un conjunto de preguntas a tratar como candidatas "on the fly" y a partir de ahí determinar su similaridad.

Respuesta

Esta es una gran observación que abre muchas puertas a distintos diseños de RSs. Efectivamente, el diseño propuesto por este trabajo considera un proceso fuera de línea que realiza comparaciones de preguntas "todas contra todas". Si bien este enfoque no es óptimo computacionalmente, se tomó como base para el desafío de la creación de una nueva arquitectura distribuida. Los experimentos, por otro lado, se realizaron tomando muestras aleatorias con significancia estadística.

Con respecto a la consideración de "candidate selection", se pueden agregar las siguientes observaciones:

- No se consideró ningún proceso especial para la reducción de la cantidad de cálculos de similaridad para cada ejecución de experimentos o para la propuesta del cálculo fuera de línea, lo cual podría haber sido un buen enfoque para la optimización de recursos computacionales. Se dejó esa tarea a cada una de las medidas de similaridad: las comparaciones que arrojaron similaridad 0, fueron descartadas para el proceso de clustering y etiquetado. Se generaron matrices más dispersas para métodos basados en term-frequency (TF y TF-IDF), en comparación con términos basados en representaciones vectoriales (FastText y Word2Vec) o taxonomías (Semantic Distance).

- Por otro lado, se consideraron varios enfoques para el cálculo de similaridad en tiempo de ejecución (agregar una nueva pregunta). La idea de la presentación de los mismos fue intentar resolver el problema de "cold star". Se observa que sería posible utilizar estos métodos para realizar un proceso de "candidate selectio" pero en tiempo real. Por ejemplo, es posible utilizar un método KNN entre las representaciones vectoriales para luego utilizar la respuesta y calcular la similaridad utilizando EQuALs, pero solo entre las preguntas pertenecientes a ese subconjunto.

Comentario 5

Uno de los problemas más comentados en lo que se refiere a sistemas que utilizan pares de preguntas de Quora es que muchas de las preguntas que parecen similares, no lo son. Básicamente, la literatura en lo que se refiere al state-of-the-art menciona que pares de preguntas que generan altos valores de similaridad (independientemente de la medida usada para establecer similaridad) pueden referirse a conceptos que en realidad no están relacionados (i.e., el objetivo de las preguntas no es el mismo). Me pregunto si se consideró como puede afectar este problema los distintos casos de uso presentados en la Sección 4.5 de este manuscrito.

Respuesta

Decir que no se considero, pero que se podría resolver de cierta forma.

Comentario 6

En la Tabla 3 se muestran matrices de confusión, sería más interesante (y simple visualmente) incluir simplemente Accuracy (Exactitud), False Positives y False Negatives. Error es redundante, ya que es Total – Exactitud, con lo cual va a ayudar a la legibilidad de la tabla de resultados remover esa columna.

Respuesta

Se presenta una nueva tabla, siguiendo las recomendaciones:

Tabla 1: Matrices de confusión para los cinco algoritmos de medidas de similaridad.

			Predicho		**Exactitud**	**Error**
			0	**1**		
TF	**Real**	**0**	0.4355	0.1953	0.6776	0.3224
		1	0.1271	0.2421		
TF/IDF	**Real**	**0**	0.4477	0.1831	0.6685	0.3315
		1	0.1484	0.2208		
Word2Vec	**Real**	**0**	0.4343	0.1965	0.6788	0.3212
		1	0.1247	0.2445		
FastText	**Real**	**0**	0.5033	0.1275	0.6725	0.3275
		1	0.2	0.1692		
Semantic Distance	**Real**	**0**	0.4877	0.1431	**0.6797**	**0.3203**
		1	0.1772	0.192		

Comentario 7

Es también importante mencionar que las medidas reportadas no se definen hasta varias páginas después. Esto es contraproducente, por lo cual sugiero comenzar el capítulo definiendo claramente medidas de evaluación que después se utilizan en el resto de la narrativa.

Respuesta

Se debería mover la sección “5.6.2.1. Estructura de las matrices de confusió” al principio del “Capítulo 5. Experimentos”. De tal forma, cuando se presentan los resultados del trabajo del estado del arte, el lector puede hacer una mejor lectura de los mismos.

Comentario 8

Siguiendo con los experimentos, proporcionaría contexto a los resultados describir con detalles antes de mostrar la tabla la distribución de pares de preguntas que se utilizan en la evaluación. Es decir: cuántos pares de preguntas están asociadas al rótulo 0 y cuantos al rótulo 1, ya eso impacta la valoración de "Exactitud".

Respuesta

En la sección "5.1.1. Medidas de rendimiento y error" se detalla la distribución de pares de preguntas de clase 0 y 1 como "$36,9\,\%$ pares de preguntas son clase 1 y el $63,1\,\%$ restante es clase 0". Podría haber sido conveniente ubicar esta aclaración antes de la Tabla 3.

Nota: cada uno de las ejecuciones realizadas en los experimentos de este trabajo utilizan muestras aleatorias que garantizan que la proporción de preguntas de clase 1 esté entre $35\,\%$y $65\,\%$ del total de preguntas del subconjuntos, para dar significancia estadística y poseer una variabilidad de datos que derive en resultados confiables.

Comentario 9

Lo más importante, a los resultados presentados en la Capítulo 5 (y en el análisis descrito en el Capítulo 6) les falta un test de significancia estadística, de lo contrario no es posible establecer la veracidad de las conclusiones. Por ejemplo, en la tabla 3 se señala semantic distance como la mejor medida de similaridad, sin embargo, sin test de significancia estadística no es posible establecer que es realmente la medida más efectiva comparada con las demás. En la tabla 25, se indica que 30 es el valor de K que produce el menor error. Sin embargo, la falta de test de estadística dificulta la posibilidad de concluir definitivamente que por ejemplo K=35 o k=50 no serían alternativas similarmente factibles. En la Sección 6.3.1, se presenta un análisis de varianza del método propuesto. Claramente esto responde en parte a mi pregunta en lo referido a la veracidad de las conclusiones, pero no se menciona sino hasta muy tarde en el manuscrito. Creo que es imperativo que se defina claramente el objetivo de los experimentos, el dataset y las medidas de evaluación y análisis al inicio del Capítulo 5. De esta manera el lector podrá seguir la descripción de los resultados. También sugiero

que se mencione explícitamente en los Capítulos 5 y 6 los resultados que son significativos a medida que se introducen, no al finalizar el análisis.

Respuesta

Algunas observaciones sobre esta oportunidad de mejora:

- La fila resaltada en la Tabla 30, que indica que con un valor de $k = 30$ se obtiene el mejor error, es solo un agregado visual para ayudar a la interpretación de resultados en esa fase de la experimentación. Lo mismo sucede para las tablas 25-30. Adicionalmente, se provee la Figura 15, que permite visualizar cómo los valores de error decrecen a medida que el número de clusters aumenta. Lo cual sugiere, con el fin de agregar valor contextual (sin aportar ventajas comparativas contra los métodos del estado del arte), que el método EQuAL se comporta mejor con valores de k grandes, hasta cierto valor que se considera "óptimo" teniendo en cuenta el método del codo.

- La idea del tesista fue introducir la interpretación de resultados a "bajo nivel" a medida que se introducen los experimentos en el trabajo de tesis, e intentar resumirlos en un nivel más alto cuando termina la sección.

- En análisis de varianza realizado en la sección "6.3.1. Análisis de varianza del método propuesto" utiliza los resultados de la sección "6.2. Análisis del método propuesto y algoritmos del estado del arte". Por lo cual, no sería posible ubicar los mismos de una forma más temprana en el trabajo.

Cambios propuestos para la mejora de la lectura del trabajo:

- Agregar un test de significancia estadística para la tabla 3 (estado del arte).

- Definir los objetivos de los experimentos.

Comentario 10

Sería interesante expresar consideraciones en lo que se refiere a mitigación de errores en el Capítulo 6. Es decir, dado a que los umbrales de similaridades pueden generar errores (que es normal cuando umbrales son requeridos), sería interesante que el candidato mencionara como estos umbrales pueden afectar la subsecuente recomendación.

Respuesta

Es posible profundizar como, conceptualmente, los umbrales afectan la subsecuente recomendación. Se considera agregar información a la sección "5.6.2.2. Construcción y elección del umbral correcto":

> La elección del mejor umbral se realiza eligiendo valores en el intervalo $(0, 1)$ y evaluando cual de ellos conlleva a un mejor rendimiento, es decir, que los valores calculados a partir del umbral coincidan, en una mayor medida, con el valor real proveniente de la muestra de datos. Los valores de umbral tienen impacto en los falsos negativos y falsos positivos. Un valor de umbral alto, reducirá la cantidad de falsos positivos, ya que menos pares de preguntas van a ser consideradas como iguales (rótulo 1). Además, la cantidad de falsos negativos aumentará, ya que el proceso será restrictivo en cuanto a clasificar una instancia como positiva. Por el otro lado, sucederá todo lo contrario en cuanto se reduzca el valor de umbral (Fernández et al., 2018). Por este motivo, es necesario realizar un proceso iterativo para identificar el valor de umbral óptimo $t*$. Dos aspectos en tener en cuenta para este procedimiento:
>
> - El valor $t*$ debe minimizar el costo. En nuestro caso, por cada uno de los valores de t que se configuren como parámetro, la salida será una matriz de confusión. El valor de umbral t que produzca la matriz con el menor error, será el elegido.
> - Es posible utilizar diferentes intervalos entre los valores de t de entrada. Por ejemplo, si tomamos un intervalo de 0,05, obtendremos 19 valores de t ($[0{,}05, 0{,}1, 0{,}15, ..., 0{,}90, 0{,}95]$). En cambio, si el intervalo es de 0,01 obtendremos 100 valores potenciales de t ($[0{,}01, 0{,}02, 0{,}0{,}3, ..., 0{,}98, 0{,}99]$). Claramente, mientras más chico sea el intervalo de elección de umbral, tendremos más oportunidades de obtener un error más bajo, a cambio de mayor costo computacional.

Bibliografía

Cormen, Thomas H; Leiserson, Charles E; Rivest, Ronald L y Stein, Clifford (2009). *Introduction to algorithms*. MIT press.

Erk, Katrin y Priese, Lutz (2008). *Theoretische Informatik: Eine umfassende Einführung*. Springer-Verlag.

Fernández, Alberto; García, Salvador; Galar, Mikel; Prati, Ronaldo C; Krawczyk, Bartosz y Herrera, Francisco (2018). *Learning from imbalanced data sets*, tomo 10. Springer.

Goldberg, David; Nichols, David; Oki, Brian M y Terry, Douglas (1992). «Using collaborative filtering to weave an information tapestry». *Communications of the ACM*, **35(12)**, pp. 61–70.

Gonzalez, Alejandro; Flury, Carlos; Ferrari, Franco; Guereta, Guadalupe; Valoni, Mercedes; Diez, Santiago; Pera, Sole; Madrazo Azpiazu, Ion y Leale, Guillermo (2017). «Comparative Analysis on Text Distance Measures Applied to Community Question Answering Data». En: *5to Congreso Nacional de Ingeniería Informática / Sistemas de Información CONAIISI 2017, Santa Fe, Argentina*, tomo 1, pp. 99–106.

Hill, Will; Stead, Larry; Rosenstein, Mark y Furnas, George (1995). «Recommending and evaluating choices in a virtual community of use». En: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 194–201.

Jain, Anil K (2010). «Data clustering: 50 years beyond K-means». *Pattern recognition letters*, **31(8)**, pp. 651–666.

Nigam, Kamal; McCallum, Andrew Kachites; Thrun, Sebastian y Mitchell, Tom (2000). «Text classification from labeled and unlabeled documents using EM». *Machine learning*, **39(2)**, pp. 103–134.

Peña, Daniel (2013). *Análisis de datos multivariantes*. McGraw-Hill España.

Resnick, Paul; Iacovou, Neophytos; Suchak, Mitesh; Bergstrom, Peter y Riedl, John (1994). «Grouplens: An open architecture for collaborative filtering of netnews». En: *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186.

Rich, Elaine (1979). «User modeling via stereotypes». *Cognitive science*, **3(4)**, pp. 329–354.

Sarwar, Badrul; Karypis, George; Konstan, Joseph y Riedl, John (2001). «Item-based collaborative filtering recommendation algorithms». En: *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295.

Sarwar, Badrul; Karypis, George; Konstan, Joseph y Riedl, John (2002). «Incremental singular value decomposition algorithms for highly scalable recommender systems». En: *Fifth international conference on computer and information science*, tomo 1, pp. 27–8. Citeseer.

Schafer, J Ben; Frankowski, Dan; Herlocker, Jon y Sen, Shilad (2007). «Collaborative filtering recommender systems». En: *The adaptive web*, pp. 291–324. Springer.

Shardanand, Upendra y Maes, Pattie (1995). «Social information filtering: algorithms for automating "word of mouth"». En: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 210–217.

Xu, Rui y Wunsch, Don (2008). *Clustering*, tomo 10. John Wiley & Sons.