Análisis comparativo de la aplicación de monogramas y digramas en la clasificación de documentos

M.Sc. Víctor Manuel Cornejo Aparicio[1,2] Jenny Copara Zea[2]
vcornejo5@hotmail.com, jen_copara@hotmail.com

[1]Universidad Nacional San Agustín de Arequipa
Av. Independencia S/N - Cercado
[2]Universidad Alas Peruanas Filial Arequipa
Urb. Daniel Alcides Carrión G-14 Dist. José Luis Bustamante y Rivero
Arequipa – Perú

Resumen: *En este artículo se presenta el resumen de la investigación en el área de procesamiento de lenguaje natural (Natural Language Processing), la aplicación de modelos de espacios de palabras (Word Space Model) en la clasificación automática supervisada de documentos, empleando monogramas y digramas, donde el propósito fundamental es la comparación de la efectividad de la clasificación de estos ngramas.*

Abstract: *This article is a summary of research in the area of natural language processing (Natural Language Processing), the application of space models of words (Word Space Model) in supervised automatic classification of documents, using monograms, bigrams, where the main purpose is to compare the effectiveness of the classification of these ngramas.*

Palabras clave: Procesamiento de Lenguaje Natural, Modelo de Espacio de Palabras, Clasificación de Documentos, nGramas.

1. Introducción

En las diversas instituciones elaboran documentos de diversa índole. Éstos son redactados en forma regular por personas bien definidas, por su cargo o responsabilidad, además de tener formatos de redacción estandarizados. Dicho de otra forma, las personas ocupan cargos, y en ellos redactan oficios, cartas, memorandos, informes, etc. Todo esto da origen a un conjunto de características que de alguna manera configuran un estereotipo, sumado al factor de que cada persona tiene un vocabulario limitado, y es recurrente en el uso de diversas palabras.

Todos los aspectos anteriormente descritos constituyen un conjunto de patrones que son susceptibles de emplear para el reconocimiento de los tipos de documentos que generan. En el procesamiento de lenguaje natural, existe la técnica del modelamiento del espacio de palabras, el mismo que trata de la asociación de los diversos vocablos a los documentos que los contiene, lo cual constituye, en suma, un patrón de clasificación. En este contexto, surge la duda de que si un vocablo está directamente asociado en algún grado de importancia con un tipo de documento, y si la asociación de dos vocablos que constituyen mayor cantidad de datos, lo que daría a entender una mayor cantidad de información y que, por consiguiente, podría aportar mayor precisión en el tratamiento de la clasificación de documentos, este es el problema que trata de abordar el presente artículo, que básicamente tratará de demostrar qué tan bueno es tratar de efectuar trabajos de clasificación empleando monogramas y digramas de palabras lematizadas.

2. Trabajos Previos

Un equipo constituido en la Universidad Europea de Madrid – CEES, que trabaja el área de procesamiento de lenguaje natural, ha creado una herramienta denominada CADOC: "herramienta de clasificación automática de documentos" [Gómez, 2003]. Su proceso está enfocado en la extracción del texto de los documentos, el indexado de los mismos y su posterior tratamiento con el weka, para posteriormente efectuar la clasificación. Su herramienta permite hacer una clasificación definida por el usuario.

Arturo Montejo Ráez, en el Departamento de Informática de la Universidad de Jaén de España, trabajo su tesis doctoral titulada "Clasificación Automática de Textos en el Dominio de la Física de Altas Energías" [Montejo 2005], quien desarrolló principalmente sus investigaciones en el Laboratorio Europeo para la Investigación Nuclear (CERN), su planteamiento sobre clasificación de documentos lo trabajó en base a tres disciplinas: Recuperación de Información (RI), Procesamiento del Lenguaje Natural (PLN) y algoritmos de Aprendizaje Automático (Machine Learning - ML), empleó la introducción de información bibliográfica como factor importante en los resultados del proceso de clasificación

Peláez J.I. y Sánchez P. del Dpto. Lenguajes y Ciencias de la Computación, E.T.S.I. Informática. Campus de Teatinos. Universidad de Málaga España, conjuntamente que con La Red D. del Dpto. de Informática de la Universidad Nacional del Nordeste - Corrientes. Argentina, han desarrollado el trabajo denominado "Un Clasificador de Texto Por Aprendizaje" [Peláez 2002], quienes trabajan en el área de telemedicina donde buscan elaborar un clasificador estomatológico de pacientes para según ésta, sean derivados a áreas especializadas de acuerdo con la categorización definida. Su aprendizaje está basándose en los prediagnósticos establecidos por un profesional médico no especializado, y un diccionario de términos estomatológicos, es capaz de clasificar nuevos prediagnósticos en las especialidades.

3. Clasificación automática de documentos

3.1. Premisas de la investigación

En el presente trabajo, se inicia partiendo de un conjunto de premisas, las mismas que justificarán las acciones desarrolladas y cuyos mecanismos y resultados se presentan.

Premisa 1: Los documentos tienen una naturaleza y estructura, los mismos que a su vez están constituidos por textos que son un conjunto de vocablos que son regularmente empleados en documentos de similar categoría.

Premisa 2: Los vocablos individualmente constituyen información, y éstos a la vez, que se asocian entre sí, incrementan el volumen de información, la misma que podría caracterizar en mejor manera a los documentos que los contengan.

Premisa 3: Cuando se emplean más de un vocablo, en un proceso de clasificación automática (n-gramas), puede darse el caso, que una conjunción de vocablos (A, B), pueda presentarse como (B, A) en el mismo documento o uno similar del mismo tipo, lo cual en términos prácticos, constituiría una dispersión de las frecuencias asociadas a la categoría definida, para lo cual, dado el caso se debería indexar horizontalmente los vocablos, y de esta forma evitar la dispersión de las frecuencias.

Premisa 4: Al constituirse los vocablos asociados de uno, dos o más, éstos se deberán catalogar asociados al tipo de documento que les dio origen. Una vez constituida la asociación y elaborado la concentración de frecuencias, estos vocablos se asumirán como únicos a efectos de desarrollar los cálculos requeridos para la determinación de las proximidades entre los vocablos y el tipo de documento asociado.

3.2. Proceso de clasificación

El proceso de clasificación expresado en una forma muy breve, se presenta en la figura 1, la misma que consta de dos etapas plenamente diferenciadas, la etapa de entrenamiento y la etapa de clasificación.

Debido a que el propósito del presente artículo es presentar la comparación de usar monogramas o digramas en el proceso de clasificación automática, detallaremos únicamente aquellos aspectos relevantes a dicho proceso.

Para iniciar el proceso de entrenamiento, es necesario contar con un número determinado de documentos preclasificados, esto para que en el entrenamiento se puedan crear los nGramas de forma asociada al tipo definido.

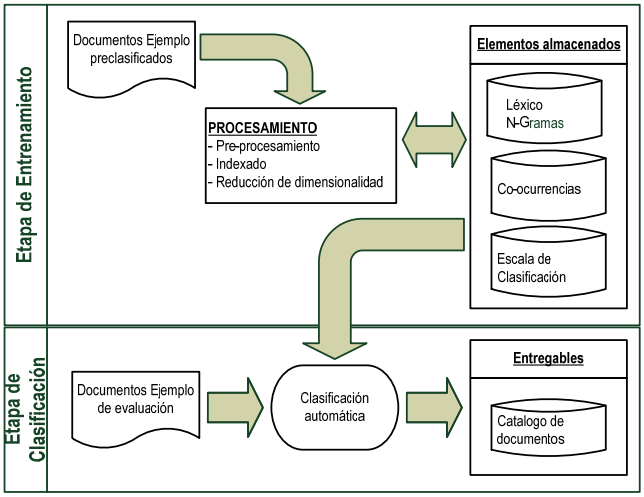

Figura 1. Esquema del proceso de clasificación.

El procesamiento consta de tres subetapas, que son: Preprocesamiento, Indexado y Reducción dimensional, en este contexto, durante el pre-procesamiento, es necesario limpiar el texto de forma tal que se pueda emplear un texto limpio de enlaces, caracteres especiales, números, símbolos u otros elementos que no aporten mayor información. Una vez logrado este aspecto, se procede a efectuar un lematizado del texto, que en suma nos entrega un texto listo para ser empleado en las tareas siguientes. Esto se puede apreciar en la figura 2 que presentamos a continuación:

Figura 2. Texto pre-procesado.

El indexado de la información se da en dos términos, el primero se da cuando se trata de asociaciones de vocablos de más de un término (digrama o superior) en sentido horizontal, esto debido a que las asociaciones (A, B) y (B, A), es la misma conjugación, mantenerlas separadas, constituiría una dispersión de la información, y en segundo plano es un indexado vertical donde se ordena en forma creciente los términos con una jerarquía definida de la forma siguiente: Tipo de documento, raíz 1, raíz 2, …raíz n. Cabe recordar que el presente trabajo solo presenta el caso de los monogramas y digramas, los mismos que se muestran en los cuadros 1 y 2.

Doc	Raiz	Frec.
Carta	arequip	3
Carta	juni	1
Carta	señor	1
Carta	president	1
Carta	colegi	2
Carta	ingenier	3
Carta	per	2
…	…	…
Carta	autorizac	1
Carta	expuest	1
Carta	atent	1
Carta	ing	1

Doc	Raiz1	Raiz2	Frec.
Carta	arequip	juni	1
Carta	juni	señor	1
Carta	president	señor	1
Carta	colegi	president	1
Carta	colegi	ingenier	2
Carta	ingenier	per	2
Carta	consej	per	2
…	…	…	…
Carta	atent	expuest	1
Carta	atent	ing	1
Carta	ing	juan	1
Carta	dni	garci	1

Cuadro 1. Tabla de ejemplo de monogramas y digramas indexado horizontalmente.

Doc	Raiz	Frec.
Carta	acces	1
Carta	acertad	1
Carta	Año	1
Carta	arequip	3
Carta	Atent	1
Carta	autorizac	1
Carta	canch	1
Carta	Cãs	2
…	…	…
Carta	sistem	1
Carta	solicit	2
Carta	Total	1
Carta	Urb	1
Carta	Uso	1

Doc	Raiz1	Raiz2	Frec.
Carta	acces	local	1
Carta	acces	solicit	1
Carta	acertad	esper	1
Carta	acertad	favorab	1
Carta	año	part	1
Carta	año	present	1
Carta	arequip	cuent	1
Carta	arequip	departament	2
Carta	arequip	juni	1
Carta	arequip	solicit	1
…	…	…	…
Carta	present	respet	1
Carta	president	señor	1
Carta	ros	urb	1

Cuadro 2. Tabla de ejemplo de monogramas y digramas indexado verticalmente.

Posteriormente, se efectúa la reducción dimensional, la misma que reducirá notablemente el número de términos con los cuales se trabajará la matriz de coocurrencia, esto es efectuado básicamente para no sobrecargar los algoritmos al aplicar la clasificación y reducir su tiempo de ejecución.

El proceso de entrenamiento en su núcleo central se trabaja con un algoritmo que contenga tres parámetros básicos: El texto lematizado a procesar, el nGrama que se desea construir, y el identificador del tipo de documento al que pertenece el texto, todo ello se elabora en los pasos siguientes:

```
Procedimiento EntrenarNGrama
Parametros: Texto     'Texto pre-procesado y lematizado
   nGrama  'Tipo de nGrama a Procesar [1] Monograma, [2] Digrama
   IdDocTipo 'Tipo de documento al que corresponde el entrenamiento
   Raiz = ExtraerPalabra(Texto)
   Mientras Raiz <> ""
      IdRaiz =MatricularRaiz(Raiz)
      InsertarRaiz(Idraiz, vRaiz)
      vRaizOrdenado = OrdenarVector( vRaiz)
      MatricularNGrama(IdDocTipo, vRaizOrdenado)
      Raiz = ExtraerPalabra(Texto)
   Fin Mientras
Fin Procedimiento
```

4. Experimentos y Resultados

Uno de los experimentos trabajados en la investigación respecto al tema planteado se efectuó con un conjunto de documentos definidos en el siguiente cuadro:

Tipo de Documento	Cantidad	Muestra
Oficio	180	40
Carta	342	45
Solicitud	188	41
Memorando	179	40
Contrato	177	40
Informe	187	41
Recibo	919	49

Cuadro 3. Tabla de cantidad de tipos de documentos empleados.

Para el proceso de clasificación, se empleará una muestra aleatoria, la misma que se determinó en base a la siguiente fórmula estadística, y cuyos resultados se muestran en el cuadro 3:

$$n = \frac{N}{1 + \frac{e^2(N-1)}{z^2 pq}}$$

Donde:

n: Tamaño de la muestra que deseamos obtener

N: Tamaño conocido de la población

e: 0.05

z: 1.65

p: 5%

q: 95%

De la muestra seleccionada, según los tipos de documentos, se seleccionaron cinco motivos de clasificación. Se efectuó esta acción, pues en algún momento se trató de establecer la concordancia con la estructura de los documentos. Los motivos que se seleccionaron fueron

1. Cartas de presentación.
2. Ascensos.
3. Contratos de prestación de servicios.
4. Procedimientos administrativos de grado.
5. Certámenes Académicos.

Efectuado el proceso de entrenamiento, y luego de construir el corpus de monogramas y digramas correspondientes a los modelos de documentos con los motivos establecidos, se procedió a efectuar el proceso de clasificación empleando para ello los monogramas y digramas, en un primer momento empleando estos ngramas de forma original y en un segundo tiempo aplicando una reducción dimensional.

Los resultados obtenidos en el caso de los monogramas sin la aplicación de reducción dimensional, se muestran en el cuadro número 4. En esa tabla, se puede observar que el proceso de clasificación es aceptable, pero existe un grado de precisión del 83%, lo cual nos indica que no es del todo aplicable para casos que requiera un nivel de confiabilidad superior.

Motivo	1		2		3		4		5		Total
	F	%	F	%	F	%	F	%	F	%	
1	48	80%	7	12%	0	0%	1	2%	4	7%	60
2	5	8%	45	75%	0	0%	7	12%	3	5%	60
3	0	0%	1	2%	55	92%	3	5%	1	2%	60

Motivo	1		2		3		4		5		Total
	F	%	F	%	F	%	F	%	F	%	
4	3	5%	0	0%	0	0%	51	85%	6	10%	60
5	4	7%	1	2%	0	0%	5	9%	46	82%	56
										Total	296

Cuadro 4. Tabla de cantidad de tipos de documentos empleados.

Los resultados obtenidos en el caso de los digramas sin la aplicación de reducción dimensional, se muestran en el cuadro número 5. En éste se presentan resultados nada alentadores para el empleo de digramas como técnica de clasificación, pues solo alcanza un nivel de precisión promedio del orden del 78%, lo cual podría juzgarse erróneamente de forma apresurada como una técnica no confiable.

Motivo	1		2		3		4		5		Total
	F	%	F	%	F	%	F	%	F	%	
1	45	75%	8	13%	0	0%	1	2%	6	10%	60
2	8	13%	40	67%	0	0%	3	5%	9	15%	60
3	0	0%	6	10%	52	87%	2	3%	0	0%	60
4	5	8%	2	3%	0	0%	43	72%	10	17%	60
5	3	5%	1	2%	0	0%	3	5%	49	88%	56
										Total	296

Cuadro 5. Tabla de cantidad de tipos de documentos empleados.

Los resultados obtenidos en el caso de los monogramas con la aplicación de reducción dimensional, reduce notablemente la precisión de los monogramas, esto con un nivel de precisión promedio del orden del 71%, lo cual se puede apreciar en el cuadro número 6 que se presenta a continuación.

Motivo	1		2		3		4		5		Total
	F	%	F	%	F	%	F	%	F	%	
1	43	72%	9	15%	0	0%	3	5%	5	8%	60
2	12	20%	37	62%	0	0%	6	10%	5	8%	60
3	0	0%	1	2%	53	88%	6	10%	0	0%	60
4	14	23%	2	3%	0	0%	37	62%	7	12%	60
5	6	11%	4	7%	0	0%	7	13%	39	70%	56
										Total	296

Cuadro 6. Tabla de cantidad de tipos de documentos empleados.

Los resultados obtenidos en el caso de los digramas con la aplicación de reducción dimensional mejoran notablemente la precisión del proceso de clasificación, alcanzando un promedio del orden del 95%. Dichos resultados se evidencian en el cuadro número 7, el mismo que se presenta a continuación.

Motivo	1		2		3		4		5		Total
	F	%	F	%	F	%	F	%	F	%	
1	53	88%	4	7%	0	0%	1	2%	2	3%	60
2	1	2%	57	95%	0	0%	1	2%	1	2%	60
3	0	0%	0	0%	59	98%	1	2%	0	0%	60
4	1	2%	0	0%	0	0%	58	97%	1	2%	60
5	1	2%	0	0%	0	0%	1	2%	54	96%	56
										Total	296

Cuadro 7. Tabla de cantidad de tipos de documentos empleados.

5. Conclusiones

Luego de efectuadas las pruebas experimentales donde se entrenaron y luego clasificaron los documentos en un ambiente controlado y supervisado, se puede decir que a medida que se incrementa la diversidad de tipos de documento con estructuras diversas o no muy bien definidas, los monogramas arrojan mejores resultados que cuando no se aplica la reducción dimensional. Esto también se puede observar tomando como medida de comparación el tamaño del corpus generado, y puede decirse que a medida que el corpus de documentos crece los monogramas son más efectivos. Pero con documentos bien estructurados, y con una reducción dimensional, los digramas mejoran su rendimiento y precisión.

Podría parecer que la estructura de los documentos es irrelevante, puesto que al preprocesar los textos contenidos, esta estructura se pierde, lo cual no es del todo correcto. En formato, la estructura de los párrafos puede perderse, pero la secuencia de los vocablos relevantes permanece, lo cual se pudo evidenciar en tipos con estructura muy rígida, como es el caso de los contratos, donde en todos los casos su clasificación fue efectiva. En los oficios, esto sucedió en el noventa por ciento, y así disminuye en cuanto la estructura se hace más diversa como es el caso de las cartas

Se puede sugerir trabajos futuros en el orden de determinar el impacto de la estructura de los documentos en el proceso de clasificación, así como la generación personalizada de corpus por autor, para ver la autenticidad de los documentos. También sería pertinente establecer umbrales de reducción dimensional por ganancia de información por cada motivo.

Para artículos futuros se está experimentando con corpus Reuters21578-Apte-90Cat y Reuters21578-Apte-115Cat, para repetir el análisis exento de estructura, y con ello concretar un juicio de mayor precisión.

Referencias bibliográficas

[Montejo, 2010] Montejo A., Perea J.M., Martín M. y Ureña A., "Uso de la detección de bigramas para categorización de texto en un dominio científico", Revista Procesamiento de Lenguaje Natural, No 44 (2010).

[Cavnar 1994] Cavnar W. and Trenkle J., "N-Gram-Based Text Categorization", 3rd Annual Symposium on Document Analysis and Information Retrieval.

[Gómez, 2003] Isidro Gómez Mompó, Jaime Lozano Muñoz, Diego Martínez Salazar, Luis Muñoz Góngora, Diego Ramírez Adrados, "CADOC: herramienta de clasificación automática de documentos", Universidad Europea de Madrid – CEES, Disponible en:

http://www.esp.uem.es/jmgomez/plenum/plenum3/03.pdf , Junio 2003.

[Montejo 2005] Arturo Montejo Ráez, Tesis Doctoral "Automatic Text Categorization of Documents in the High Energy Physic Domain", Departamento de Informática de la Universidad de Jaén de España, Diciembre del 2005.

[Peláez 2002] Peláez J.I. y Sánchez P. "Un Clasificador de Texto Por Aprendizaje", Revista Inteligencia Artificial, Vol 6, No 15 (2002), disponible en:

http://aepia.lcc.uma.es/index.php/ia/article/view/756.

[Magnus 2006] Magnus Sahlgren, Tesis Doctoral “The Word-Space Model - Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces”, Stockholm University, 2006.

[Tejada 2009] Javier Tejada Cárcamo, Tesis doctoral “Construcción automática de un modelo de espacio de palabras mediante relaciones sintagmáticas y paradigmáticas”, Instituto Politécnico Nacional, Centro De Investigación En Computación, México 2009.