ANÁLISIS COMPARATIVO DE LA APLICACIÓN DE MONOGRAMAS Y DIGRAMAS EN LA CLASIFICACIÓN DE DOCUMENTOS

Comparative analysis in the application of monograms and digrams in the classification of documents

Víctor Manuel Cornejo Aparicio[1], Jenny Linet Copara Zea[2]

RESUMEN

El presente trabajo se desarrolla en el área de procesamiento de lenguaje natural (Natural LanguageProcessing), la aplicación de modelos de espacios de palabras (Word SpaceModel) 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.

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

ABSTRACT

This paper presents a summary of the research in the area of Natural Language Processing, the application of Word Space Model in the supervised automatic classification of documents, using monograms and bigrams, where the primary purpose is to compare the effectiveness of the classification of these ngramas.

Keywords*: Processing in natural language, Word space model, Classification of documents, nGrams.*

[1]Dr(c) en Ciencias de Computación. Docente en la Universidad Alas Peruanas Filial Arequipa y Universidad Nacional de San Agustín. E-mail: vcornejo5@hotmail.com
[2] Bachiller en Ingeniería de Sistemas e Informática de la Universidad Alas Peruanas Filial Arequipa. E-mail: jennycopara@gmail.com

INTRODUCCION

En las diversas instituciones se elaboran documentos de diferente índole, estos son regularmente redactados por personas que poseen un adecuado manejo de la redacción, sea por su cargo o responsabilidad, además de tener formatos de redacción estandarizados. Sin embargo, en algunas situaciones, no se presenta una situación comunicativa que plantee exactamente lo que se pretende desarrollar. Por lo tanto, todo ello 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 al redactar sus documentos.

Por lo anteriormente descrito, se ha venido a constituir un conjunto de patrones que son susceptibles de emplear para el reconocimiento de los tipos de documentos que se 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. Desde 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, esta es la problemática que trata de abordar el presente artículo, que básicamente tratará de demostrar que tan bueno es tratar de efectuar trabajos de clasificación empleando monogramas y diagramas de palabras lematizadas.

MATERIAL Y MÉTODO

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 diagramas 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 N-Gramas de forma asociada al tipo definido.

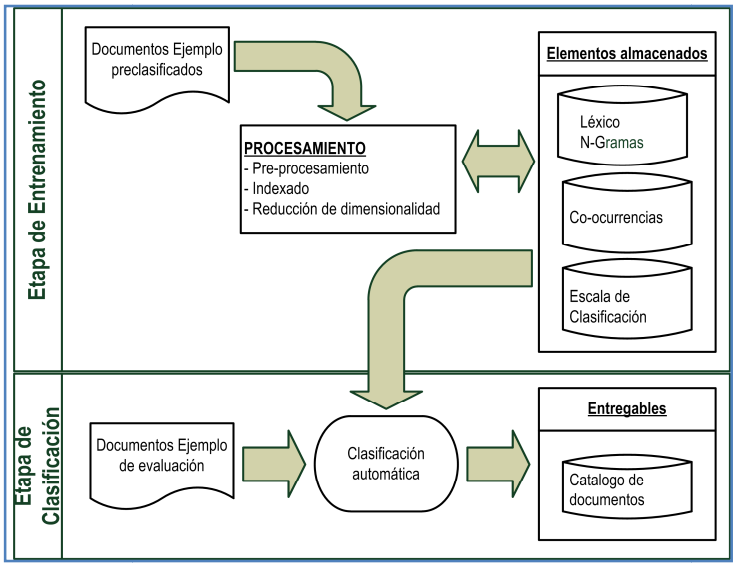

Figura 1. Esquema del proceso de clasificación

El procesamiento consta de tres sub etapas, que son: Pre-procesamiento, Indexado y Reducción dimensional. Desde esta perspectiva, 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:

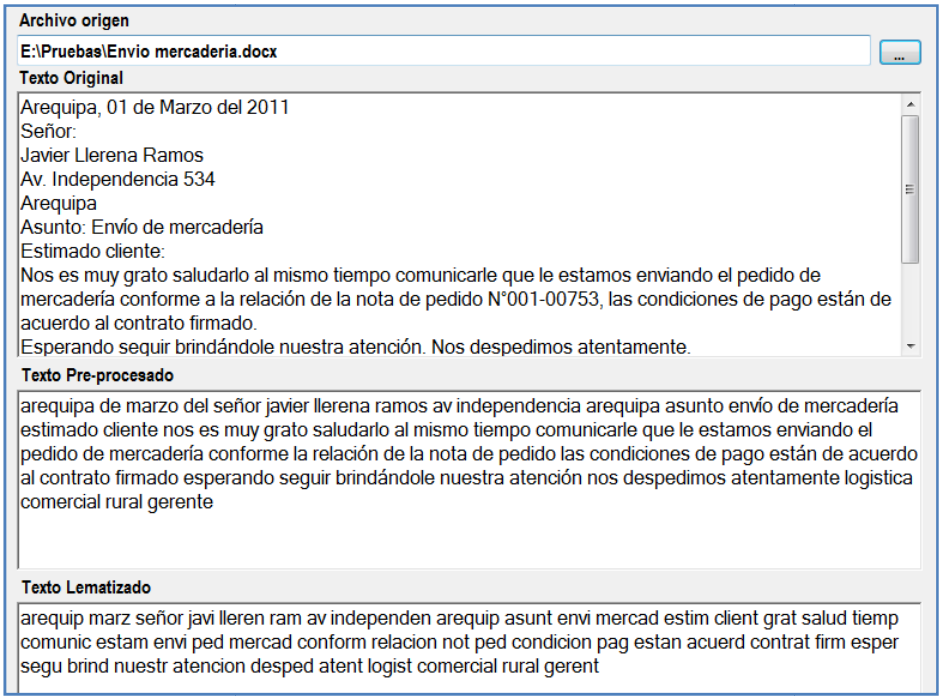

Figura 2. Texto pre-procesado

El indexado de la información se da en dos términos: el primero se produce cuando se trata de asociaciones de vocablos de más de un término (diagrama o superior) en sentido horizontal, porque si las asociaciones (A, B) y (B, A), las cuales comprenden la misma conjugación, se mantuvieran separadas, constituiría una dispersión de la información; en segundo plano, se trata de 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 las Tablas1 y 2.

Tabla 1.

Ejemplo de monogramas y digramas indexado horizontalmente

Doc	Raiz	Frec.	Doc	Raiz1	Raiz2	Frec.
Carta	arequip	3	Carta	arequip	marz	1
Carta	marz	1	Carta	marz	señor	4
Carta	señor	3	Carta	javi	señor	2
Carta	javi	1	Carta	javi	lleren	3
Carta	lleren	1	Carta	lleren	ram	2
Carta	ram	2	Carta	av	ram	2
Carta	av	1	Carta	av	independen	1
…	…	…	…	…	…	…
Carta	condic	1	Carta	condic	pag	1
Carta	pag	1	Carta	estan	pag	1
Carta	estan	1	Carta	acuerd	estan	1
Carta	acuerd	1	Carta	acuerd	contrat	1

Fuente: Elaboración propia

Tabla 2.

Ejemplo de monogramas y digramas indexado verticalmente

Doc	Raiz	Frec.	Doc	Raiz1	Raiz2	Frec.
Carta	acuerd	1	Carta	acuerd	estan	1
Carta	arequip	3	Carta	acuerd	contrat	1
Carta	av	1	Carta	arequip	marz	1
Carta	condic	1	Carta	av	ram	2
Carta	estan	1	Carta	av	independen	1
…	…	…	Carta	condic	pag	1
Carta	javi	1	Carta	estan	pag	1
Carta	lleren	1	Carta	javi	señor	2
Carta	marz	1	Carta	javi	lleren	3
Carta	pag	1	…	…	…	…
Carta	ram	2	Carta	lleren	ram	2
Carta	señor	3	Carta	marz	señor	4

Fuente: Elaboración propia

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)
MientrasRaiz<> ""
IdRaiz =MatricularRaiz(Raiz)
InsertarRaiz(Idraiz, vRaiz)
vRaizOrdenado = OrdenarVector( vRaiz)
MatricularNGrama(IdDocTipo, vRaizOrdenado)
Raiz = ExtraerPalabra(Texto)
Fin Mientras
Fin Procedimiento
```

RESULTADOS

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

Tabla 3.

Cantidad de tipos de documentos empleados

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

Fuente: Elaboración propia

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 la Tabla3:

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

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. Eventos 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 diagramas, en un primer momento empleando estos ngramas de forma original y en un segundo tiempo aplicando una reducción dimensional.

ANÁLISIS Y DISCUSIÓN

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

Tabla 4.

Cantidad de tipos de documentos empleados

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
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**

Fuente: Elaboración propia

Los resultados obtenidos en el caso de los digramas sin la aplicación de reducción dimensional, se muestran en la Tabla 5, en este 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.

Tabla 5.

Cantidad de tipos de documentos empleados

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**

Fuente: Elaboración propia

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 la Tabla6 que se presenta a continuación.

Tabla 6.

Cantidad de tipos de documentos empleados

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**

Fuente: Elaboración propia

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 la Tabla 7, el mismo que se presenta a continuación.

Tabla 7.

Cantidad de tipos de documentos empleados

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**

Fuente: Elaboración propia

CONCLUSIONES

1. Luego de efectuadas las pruebas experimentales donde se entrenaron y 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 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.

2. Podría parecer que la estructura de los documentos es irrelevante, puesto que al pre-procesar 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 sí disminuye en cuanto la estructura se hace más diversa como es el caso de las cartas

3. 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.

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

Cavnar, W.,Trenkle J.(1994), *N-Gram-Based Text Categorization*, 3rd Annual Symposium on Document Analysis and Information Retrieval.

Gómez,I., Lozano,J., Martínez,D., Muñoz, L., Ramírez,D. (2003),*"CADOC: Herramienta de clasificación automática de documentos"*, Universidad Europea de Madrid – CEES.

Montejo,A. (2005), Tesis Doctoral *"Automatic Text Categorization of Documents in the High EnergyPhysicDomain"*, Departamento de Informática de la Universidad de Jaén de España.

Peláez, J. y Sánchez, P.(2002), *"Un Clasificador de Texto Por Aprendizaje*", Revista Inteligencia Artificial, Vol 6, No 15.

Sahlgren, M. (2006), 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,J. (2009), Tesis Doctoral *"Construcción automática de un modelo de espacio de palabras mediante relaciones sintagmáticas yparadigmáticas*", Instituto Politécnico Nacional, Centro De Investigación En Computación, México.