Texto Predictivo Personalizado en Dispositivos Móviles Utilizando Aprendizaje Basado en Instancias

Andres R. Luzuriaga
Universidad de Palermo, Faculta de Ingeniería,
Buenos Aires, Argentina,
andresluzu@palermo.edu

and

Santiago A. Altieri
Universidad de Palermo, Facultad de Ingeniería,
Buenos Aires, Argentina,
saltieri@palermo.edu

Abstract

Mobile devices with limited interface (numeric keypad) are equipped with predictive text technology to simplify the writing of short text messages. The prediction is based on a common words dictionary, according to the device's language, and can be extended by the use over time. The solution presented in this paper aims to complement the technology in predicting short text messages, with the inclusion of a dictionary drawn from historical messages, and personalized according to the attributes of each message. Before writing a new message, the previous messages, telephone number, date, and time, are used by the KNN (K Nearest Neighbours) instance based learning algorithm, to obtain an enriched set of words that can be entered by pressing a single key.

Keywords: Instance Based Learning, Predictive Text, Text Messages.

Resumen

Los dispositivos móviles con interfaz limitada (teclado numérico), están provistos de tecnología de texto predictivo para simplificar la escritura de mensajes cortos de texto. La predicción se basa en un diccionario de palabras comunes existente en el dispositivo, de acuerdo al idioma, y puede ser extendido con el uso a través del tiempo.
La solución presentada en este artículo, tiene por objetivo complementar la tecnología de predicción en mensajes cortos de texto, incluyendo un diccionario de palabras extraídas de mensajes históricos, y personalizado de acuerdo a los atributos de cada mensaje. Antes de escribir un nuevo mensaje, los mensajes previos, el número de teléfono, la fecha y la hora, son usados por el algoritmo KNN (Vecinos más cercanos), de aprendizaje basado en instancias, para obtener un conjunto enriquecido de palabras que pueden ser ingresadas pulsando una sola tecla.

Palabras claves: Aprendizaje Basado en Instancias, Texto Predictivo, Mensajes de Texto.

1 INTRODUCCIÓN

El Texto Predictivo es una tecnología de entrada de texto en dispositivos móviles, que permite formar palabras presionando un solo botón por cada letra en lugar de presionar repetidas veces la tecla hasta obtener la letra deseada. Esta tecnología funciona haciendo referencia a un diccionario de palabras comunes. Cuando el usuario presiona el teclado numérico, un algoritmo busca en el diccionario una lista de palabras posibles que concuerden con la combinación presionada, y muestra la opción más probable. El usuario puede confirmar la selección, continuar con la siguiente palabra o usar una tecla para ver las otras combinaciones posibles.

Ej. En el teclado típico de los teléfonos con interfaz limitada, los números corresponden a las letras como se muestra en la *Figura 1*.

Para escribir la palabra 'hola' en el modo clásico se deberá:

1. Presionar **4 (ghi)** 2 veces para obtener 'h'
2. Presionar **6 (mno)** 4 veces para obtener 'o'
3. Presionar **5 (jkl)** 3 veces para obtener 'l'
4. Presionar **2 (abc)** 1 vez para obtener 'a'

De la misma forma para obtener 'hola' con texto predictivo se deberá:

1. Presionar **4 (ghi)** 1 vez para obtener 'h'
2. Presionar **6 (mno)** 1 vez para obtener 'o'
3. Presionar **5 (jkl)** 1 vez para obtener 'l'
4. Presionar **2 (abc)** 1 vez para obtener 'a'

Figura 1

El sistema actualiza los caracteres visualizados cada vez que se presiona una tecla mostrando la palabra más probable. En este caso el texto predictivo reduce el número de botones presionados de 9 a 4.

En el lenguaje natural, los seres humanos utilizamos un conjunto acotado de palabras de acuerdo al idioma, región, contexto o interlocutor con el cual queremos establecer un dialogo. En la comunicación por mensajes de texto, esto se ve reflejado con el uso de un léxico personalizado dependiendo de las características de cada mensaje. Es decir existe un conjunto de palabras comunes directamente relacionado con los atributos de un mensaje en particular.

En la *Tabla 1* se indica la cantidad de apariciones de una palabra por número de teléfono de una muestra verdadera de 500 mensajes históricos.

	169328	529171	016536
Donde	28	17	9
Partido	0	10	0
Doc	16	0	0
Facu	6	15	12

Tabla 1

Este trabajo pretende complementar la tecnología de predicción de texto, haciendo uso de los atributos de cada mensaje enviado y recibido por el dispositivo, para proveer un diccionario generado de forma dinámica que facilite aun más la escritura de cada nuevo mensaje.

El Aprendizaje Basado en Instancias utilizado en inteligencia artificial, provee el mecanismo necesario para implementar esta solución. Cada par de mensajes enviado-recibido conforma una instancia que es almacenada en el dispositivo, para crear un historial o memoria. Al escribir un nuevo mensaje, el algoritmo *K Nearest Neighbors* [1] es

empleado para encontrar en la memoria, mensajes con características similares y construir el léxico a partir de las palabras utilizadas en los mismos.

2 MÉTODO

2.1 K Nearest Neighbors

KNN es uno de los algoritmos más simples de aprendizaje automático. Un objeto es clasificado por mayoría de votos de sus vecinos, asignando el objeto a la clase más común entre sus k vecinos más cercanos. K es un entero positivo, que determina la cantidad de vecinos utilizados para la clasificación del objeto en cuestión.

El algoritmo no es utilizado en este caso para clasificar los mensajes, sino para obtener los K mensajes vecinos de un nuevo mensaje a enviar. Los mensajes obtenidos proveen las palabras para la construcción del léxico personalizado que facilita la escritura del nuevo mensaje.

2.2 Instancias de Mensaje de Texto

Los datos proporcionados para el funcionamiento del algoritmo son obtenidos del historial de mensajes enviados y recibidos previamente por el usuario.
Cada instancia de mensaje se describe en términos de los siguientes atributos:

- Número de teléfono
- Día
- Hora
- Texto recibido
- Texto enviado

La instancia obtiene los atributos "Número de teléfono", "Día", "Hora" y "Texto Enviado" del mensaje saliente, y en caso de tratarse de una respuesta, el atributo "Texto recibido" es provisto por el mensaje entrante correspondiente.

Siendo t el número de teléfono, d el día, h la hora, r el texto recibido y e el texto enviado, los atributos de un mensaje, se define la instancia de mensaje por el vector multidimensional:

$$x_i = (t_i, d_i, h_i, r_i, e_i)$$

2.3 Cálculo de Distancia

Se utiliza el método euclidiano para el cálculo de distancia en el espacio multidimensional.

$$d(x_i + x_j) = \sqrt{(t_i - t_j)^2 + (d_i - d_j)^2 + (h_i - h_j)^2 + (r_i - r_j)^2 + (e_i - e_j)^2}$$

2.4 Representación Numérica del Texto

Es necesario realizar una transformación del texto de los atributos "Texto Recibido" y "Texto Enviado", a valores numéricos para conformar una instancia que pueda ser evaluada correctamente por el algoritmo, de acuerdo a las características de dichos atributos.

2.4.1 Diccionarios

Para la transformación del texto se utilizan dos diccionarios de palabras obtenidas al procesar los mensajes enviados y recibidos.

Cada diccionario es un conjunto de palabras con un índice numérico incremental con dominio en **N** como se ve en el ejemplo de la *Tabla 2***.**

Índice	*Palabra*
0	Donde
1	Estas
2	Hola

Tabla 2

2.4.2 Transformación

Dado un texto de n palabras, donde p_i representa la palabra con el índice i, y siendo $f(p_i)$ la función que obtiene el índice de la palabra en el diccionario. Se define la representación numérica del texto como:

$$\sum_{i=0}^{n} 2^{f(p_i)}$$

Asegurando de esta manera la unicidad de cada representación numérica de un texto.

Ej. Para el texto recibido “Hola, donde estas?”:

$p_0 = \text{"Hola"}$ $f(p_0) = 2$ $2^2 = 4$

$p_1 = \text{"Donde"}$ $f(p_1) = 0$ $2^0 = 1$

$p_2 = \text{"Estas"}$ $f(p_2) = 1$ $2^1 = 2$

$r = 4 + 1 + 2 = 7$

2.5 **Algoritmos**

2.5.1 Aprendizaje

Para cada mensaje enviado:
Actualizar los diccionarios
Crear la instancia del mensaje y agregarla a la memoria del dispositivo.

2.5.2 Predicción

Para cada nuevo mensaje:
Buscar las K instancias de mensaje más cercanas
Crear un listado de palabras con el atributo "Texto Enviado" de las K instancias encontradas.

3 RESULTADOS Y CONCLUSIONES

En pruebas de concepto se obtuvieron resultados satisfactorios en cuanto al funcionamiento de los algoritmos. Por cada mensaje enviado, los diccionarios son actualizados con tres palabras en promedio.
En el ejemplo presentado en la introducción, el número de botones presionados se reduce de 4 a 2.

El valor de K impacta en gran medida en los resultados de la predicción. Es necesario estudiar con mayor profundidad el valor óptimo de esta variable.

Es factible la implementación de una solución de este tipo en cualquier dispositivo, con la ayuda de la plataforma J2ME de Sun Microsystems.

4 REFERENCIAS

[1] Belur V. Dasarathy, editor (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, ISBN 0-8186-8930-7