ISRAEL RIBEIRO

ALGORITMO DE CONSULTA FONÉTICA SOUNDEX PARA LOJAS VIRTUAIS

Artigo como requisito parcial para a conclusão do curso de Expecialização em Desenvolvimento de Software Para Mercados Internacionais.

Orientador: Prof. Dr. Celso Yoshikazu Ishida

Curitiba - PR

2012

ISRAEL RIBEIRO

ALGORITMO DE CONSULTA FONÉTICA SOUNDEX PARA LOJAS VIRTUAIS

Artigo como requisito parcial para a conclusão do curso de Expecialização em Desenvolvimento de Software Para Mercados Internacionais

Curitiba, 17 de dezembro de 2012.

COMISSÃO EXAMINADORA

_____
Prof. Dr. Celso Yoshikazu Ishida
Universidade Federal do Paraná

_____
Prof. Dr. Ricardo Mendes Jr.
Universidade Federal do Paraná

CURITIBA

2012

DEDICATÓRIA

Dedico este trabalho a Deus, minha família, amigos e ao Dr. Celso Y. Ishida que me ajudaram e não pouparam esforços durante todo o curso.

AGRADECIMENTO

Agradeço a Deus por ter aberto as portas desta instituição, me colocado nos projetos que passei e na empresa que trabalho na presente data, por todo seu cuidado e amor por minha vida.

Aos meus familiares que me incentivaram a não desistir nos momentos difíceis.

Ao meus amigos por compartilharem suas conquistas e por confiarem em mim no momentos difíceis.

Ao meu coordenador Dr. Celso Y. Ishida que me ajudou durante todo o curso me auxiliando com seus conselhos e abrindo as portas dos projetos que participei dentro da universidade, pelo incentivo dispensado e força para continuar até o fim.

SUMÁRIO

RESUMO

No e-commerce, a procura pelo produto é o primeiro passo para a efetivação da compra. Logo, ela deve ajudar o cliente da melhor maneira, trazendo os produtos por ele desejados, mesmo que o usuário digite o nome do item pesquisado erroneamente. Afinal, o campo de busca está para o comércio eletrônico assim como o vendedor está para a loja física. Ou seja, quanto mais eficiente for o vendedor, maiores são as chances da finalização da compra. No entanto, o que se percebe no mercado de vendas on-line é uma falha na assertividade dos buscadores.

Este artigo tem por finalidade trazer uma solução de busca que mostra quantitativamente as vantagens da utilização da pesquisa fonética com o algoritmo Soundex, bem como as melhorias necessárias para aumentar a assertividade relacionada aos caracteres da língua portuguesa. Além disso, o artigo demonstra o uso do Soundex como algoritmo base para a criação de novos padrões de comparação que atendam a demanda de diferentes ramos de atividade.

Para as demonstrações, os testes que serão mencionados a seguir foram desenvolvidos a partir do uso do algoritmo Soundex voltado para o idioma inglês, e de duas adaptações da ferramenta para atender o português do Brasil e alguns segmentos do e-commerce.

Como poderá ser visto adiante, os resultados dos testes realizados com o Soundex voltado para o inglês são insatisfatórios para lojas virtuais brasileiras, independente de seu nicho de atuação. Já quando se adapta o algoritmo para o português, percebe-se uma melhora nos resultados, contudo, foi possível identificar um novo problema: pesquisas com termos que contêm números em sua composição. Realizando ajustes para contemplar números na composição dos códigos, chegou-se a um resultado satisfatório.

Ou seja, apenas a aplicação do Soundex não é suficiente. É preciso que o algoritmo esteja plenamente adaptado ao idioma português e a todas as particularidades da língua e do comportamento do consumidor brasileiro.

ABSTRACT

In e- commerce, demand for the product is the first step towards the realization of purchase. Soon, she should help the client in the best way, bringing the products desired by him , even if the user type the name of the item searched wrongly . After all, the search field is for e-commerce as well as the seller is to the physical store. That is, the more efficient the seller, the higher the chances of checkout. However, what is perceived in the online sales market is a failure in the assertiveness of search engines.

This article aims to bring a search solution that quantitatively shows the advantages of using the phonetic search with Soundex algorithm, as well as those required to increase assertiveness related to the characters of the Portuguese language improvements. Furthermore, the article demonstrates the use of the Soundex algorithm as a basis for creating new standards of comparison that meet the demand of different industries.

For those statements, the tests that will be mentioned below were developed from the use of the English language toward Soundex algorithm, and two adaptations of the tool to meet the Portuguese of Brazil and some segments of e-commerce.

As will be seen below, the results of tests done with Soundex facing the English are unsatisfactory for Brazilian online stores, regardless of your niche of expertise. However, when the algorithm adapts to the Portuguese, there is a perceived improvement in the results, however, were able to identify a new problem : research with terms that contain numbers in their composition . Making adjustments to contemplate numbers in the composition of the codes, we arrived at a satisfactory result.

That is, only the implementation of Soundex is not enough. It is necessary that the algorithm is fully adapted to the Portuguese language and all the peculiarities of the language and behavior of Brazilian consumers.

1- **INTRODUÇÃO**

O crescente aumento do comércio eletrônico no Brasil tem favorecido as empresas do setor. Devido à concorrência, é possível perceber que existe no mercado a necessidade constante de melhorar os serviços oferecidos aos clientes. Um deles é o serviço de busca de produtos.

A pesquisa é fundamental para o e-commerce. Segundo Rodrigo Schiavini sócio diretor na empresa Fbits One Stop Shop, 40% das compras on-line têm início a partir da busca. Ou seja, este serviço é bastante representativo e estratégico para o varejista virtual. A busca geral na web foi organizada pela empresa Google. Antes de sua existência era muito difícil encontrar as informações desejadas. Assim como na busca orgânica, a busca por produtos em uma loja virtual constitui um desafio constante para as lojas virtuais do Brasil, sobretudo pelo fato de que a língua portuguesa e o consumidor brasileiro possuírem características diferentes do inglês e de outros mercados.

Indiscutivelmente o dialeto regional e o nível intelectual influenciam na forma como muitas pessoas digitam, fazem pesquisas, navegam e compram na Internet, alguns indivíduos digitam o nome do produto ou da marca de forma errada. Além disso, em um país com mais de 200 milhões de habitantes, segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), é comum que um mesmo item tenha nomes diferentes em diferentes estados ou regiões. Também é preciso considerar que algumas palavras, principalmente as de origem estrangeira, têm um grau maior de complexidade de escrita, o que pode ocasionar erros ortográficos na pesquisa.

Por todos esses motivos, é preciso que as lojas virtuais tenham buscadores capazes de minimizar todos esses aspectos do e-consumidor. Para tal, o mercado de tecnologia para e-commerce desenvolveu a busca fonética, que funciona a partir de algoritmos fonéticos. Eles propõem-se a diminuir as pesquisas mal sucedidas devido a erros de digitação por meio de indexação de palavras-chave, transformando as consultas estruturadas em um comparativo por códigos.

Neste artigo, analisaremos a utilização do Soundex, um desses algoritmos fonéticos, em alguns segmentos do e-commerce brasileiro. Primeiramente, faremos uma breve apresentação do conceito e do uso dos algoritmos fonéticos. A seguir serão

expostos os problemas mais comuns na utilização da linguagem de consulta estruturada. Adiante, faremos a análise quantitativa dos algoritmos de consulta fonética e, posteriormente, analisaremos a utilização do Soundex, inclusive com a adaptação para o idioma português.

2 – DESENVOLVIMENTO

2.1 – ALGORITMO FONÉTICO

De acordo com a National Archives – US (2007), um algoritmo fonético consiste em um algoritmo para a indexação de palavras pelo som que elas representam, atribuindo a elas chaves que representam o termo. Esta ferramenta foi originalmente para a língua inglesa e, depois, alterada para outros idiomas, inclusive para o português.

Existem diversos algoritmos fonéticos utilizados atualmente para as mais variadas aplicações e segmentos do mercado. Os mais comuns são: Soundex, Daitch Mokotoff-Soundex, Kölner Phonetik, Metaphone e BuscaBR.

Neste artigo, nos concentraremos no Soundex, que foi desenvolvido, para codificar sobrenomes para uso em recenseamentos. Os Códigos Soundex são quatro sequências de caracteres compostos de uma única letra seguida de três números.

Conforme explica Mokotoff e Daitch (1985), o Daitch Mokotoff-Soundex é um refinamento do Soundex, projetado para melhor atender aos sobrenomes de origem eslava e germânica.

Já o Kölner Phonetik, como explica Kollár (2007), é bastante semelhante ao Soundex, porém mais adequado a palavras germânicas. O mesmo autor explica que o Metaphone, por sua vez, é adequado para utilização no idioma inglês, não apenas para nomes, mas também como base para muitos corretores ortográficos. O BuscaBR, segundo Braçaroto (2008), é uma adaptação do Soundex, porém voltado para a língua portuguesa.

Neste artigo, conforme já mencionado, analisaremos o Soundex como algoritmo base para as consultas; posteriormente veremos duas variações desta ferramenta, uma para atender o português brasileiro e outra para além do idioma, bem como sua interpretação de códigos numéricos originalmente ignorados, para melhorar a assertividade e atender o nicho de atuação de cada loja virtual.

Como base nos dados a seguir demonstrados, pode-se identificar qual a melhor forma de aplicar um algoritmo fonético à uma loja virtual, quais são as deficiências da utilização de tal algoritmo, assim como os problemas e as soluções tomadas para atender de forma satisfatória a assertividade das pesquisas.

2.2 – PROBLEMAS COMUNS NA UTILIZAÇÃO DA LINGUAGEM DE CONSULTA ESTRUTURADA

Conforme explica Watson (2010), Structured Query Language, ou Linguagem de Consulta Estruturada (SQL), é uma linguagem de pesquisa declarativa padrão para bancos de dados relacional (base de dados relacional). Segundo o autor, muitas das características originais do SQL foram inspiradas na álgebra relacional. Isto significa que seu comando para consulta consiste em comparar termos exatos ou contidos por meio dos comandos SELECT.

Todavia, o problema de se utilizar tal consulta pelos comandos SQL é que eles não possibilitam o tratamento de erros ortográficos ou a apresentação de sinônimos baseados no termo original procurado pelo usuário. Também não realizam a consulta pelo som da palavra: “xampu”, “champu”, “shampu”, “shampoo” são termos totalmente diferentes, porém com o mesmo som. Abaixo, o Quadro 1.1 exemplifica alguns erros comuns de digitação encontrados durante as pesquisas, enquanto no Quadro 1.2 traz alguns dos erros de fonética mais comuns entre os usuários.

Certo	Errado
nike	niky, niki, nyke, naique
melancia	melansia, melanssia
Michael	Maichael, Mychael
Jackson	Jacksom, Jeckson
marrom	marron, maron, marom

Quadro 1.1 – Erros de digitação.

Certo	Errado

nike	nique, naki
melancia	melaincia, melanssia,
Michael	Mixael, Maiqueu
Jackson	Jackzon
marrom	amarrom, marrão

Quadro 1.2 – Erros de fonética com base na pronuncia das palavras.

É preciso entender o comportamento do consumidor brasileiro nas lojas virtuais. Ao invés de efetuar uma pesquisa pelos menus da loja, o usuário prefere utilizar diretamente o campo de busca de produtos. No entanto, na maioria dos casos, esta funcionalidade do e-commerce oferece apenas uma ajuda parcial, sugerindo termos mais pesquisados, por exemplo. Esses buscadores encontram somente os produtos cujos nomes o cliente saiba com exatidão – no máximo, cobrem a marca ou alguma palavra contida em sua descrição. Não há, portanto, a cobertura de possíveis erros de digitação ou de fonética, como nos exemplos demonstrados nos quadros 1.1 e 1.2.

Os resultados da pesquisa realizada pela MarketingSherpa em 2012 indicam que 72% dos clientes de sites de comércio eletrônico não encontram resultados para pesquisas contendo erros de digitação; 56% dos sites não consideram sinônimos dos termos pesquisados e 25% dos consumidores que realizam uma busca e não encontram os itens desejados irão realizar uma pesquisa avançada e refinar suas buscas. Os compradores desejam encontrar o que procuram na primeira pesquisa realizada e quando isso não ocorre, saem desta loja virtual e tentam em um novo e-commerce.

Este comportamento do consumidor ratifica a ideia de que, quando uma loja virtual consegue fornecer aos clientes resultados de busca satisfatórios, possuem maiores chances de fechar a venda. Logo, torna-se fácil perceber a importância da eficiência, facilidade e agilidade dos buscadores no e-commerce.

Os algoritmos fonéticos, como dito anteriormente, atuam com o propósito de diminuir esses problemas, aprimorando o sistema de busca e, consequentemente, contribuindo para a melhoria das taxas de conversão das lojas virtuais.

Neste contexto, o Soundex se destaca. Trata-se como explica Mokotoff (2007) de um algoritmo fonético para indexação de nomes pelo som pronunciado em inglês. O objetivo é encontrar palavras homófonas, isto é, palavras de pronúncias iguais, transformando-as em um código. Criado em 1918 por Robert C. Russell, de Pittsburgh, Pensilvânia, nos Estados Unidos, o Soundex é voltado para a língua inglesa, compreendendo somente as 26 letras deste idioma. A primeira letra é a que compõe parte do código; as demais combinações seguem uma tabela de regras numéricas conforme a divisão proposta por Russell, baseando-se na fonética das palavras.

Russell (1922) observa:

"Há certos sons que formam o núcleo do idioma Inglês, esses sons são inadequadamente representados apenas pelas letras do alfabeto; um som às vezes pode ser representado por mais de uma letra ou combinação de letras, e uma letra ou combinação de letras podem representar dois ou mais sons." (Russell, 1922).

De acordo com Mokotoff (2007), os dígitos dos códigos do Soundex são determinados primeiramente pela exclusão das vogais A, E, I, O, U, e das consoantes H, W, Y, e pela inclusão de números, conforme a proposta de Russell. As letras repetidas consecutivamente não são consideradas. Já os dígitos das palavras que contêm menos de quatro letras são completados com zeros. É o que mostra o Quadro 2.1.

Codigo	**Letra**	**Fonética**
1	B, F, P, V	Labial
2	C, G, J, K, Q, S, X, Z	Guturais e Simbilantes
3	D, T	Dental
4	L	Palatal Fricativa
5	M, N	Nasal
6	R	Dental Fricativa

Quadro 2.1 - Exemplo dos códigos fonéticos utilizados pelo algoritmo Soundex.

O algoritmo gera uma chave respectiva para cada palavra baseando-se no som. No Quadro 2.2, podemos ver alguns exemplos de chaves geradas pelo Soundex.

Palavra	**Soundex**
Washington	W252
Wu	W000
DeSmet	D253
Jackson	J250

Quadro 2.2 – Exemplos de palavras e os respectivos códigos fonéticos

2.3 – ANÁLISE QUANTITATIVA DOS ALGORITMOS DE CONSULTA FONÉTICA

Para a análise do desempenho da busca fonética no e-commerce foram utilizadas vinte lojas virtuais ativas. Destas, cinco lojas são de suprimentos de informática, quatro de departamento, quatro de roupas e acessórios, quatro de calçados e três de suplementos alimentares.

Os termos aplicados foram coletados com base nos termos mais buscados de cada e-commerce como mostra o Quadro 2.3 e Quadro 2.4; foram selecionados a lista dos vinte termos da conta do Google Analytics[1] na loja que apresentava maior quantidade de buscas e aplicado as variações fonéticas encontradas dentre todos os termos mais buscados, sendo essa variação um erro de digitação encontrado uma ou mais vezes no; foi desconsiderado na codificação dos carateres a descrição dos produtos por conter muitas palavras que não estão diretamente ligadas ao produto, sendo considerado o nome do produto, fabricante e marca como campo codificado pelo algoritmo.

	Suplementos de Informática	**Variações fonéticas**	**Loja de Departamentos**	**Variações fonéticas**	**Roupas e acessórios**	**Variações fonéticas**
1	tela iphone	tela ifone	ar condicionado		vestido	
2	fonte samsung rv415	fonte samsumg rv415	tablet		vestido estampado	
3	fonte cce win		tv		shorts	xorts, chorts

4	fonte samsung rv410	fonte samsumg rv410	apple	aple	vestido longo	
5	hp g42		ventilador		saia	
6	tela 15.6		celular		molekinha	molequinha
7	tela 14		fogão		saia longa	
8	dv4		bebedouro		legging	legin, leging

[1]Google Analytics é uma ferramenta de monitoramento de site disponibilizada pela empresa Google

10	cooler	coler, culer	rack	raque	biquini estampado	biquine estampado
11	tela ipad	tela iped	cama box		saia jeans	saia geans
12	bateria hp g42		microondas	microndas	sapatilha	
13	rv411		iphone	ifone	cropped	croped
14	fonte para notebook	fonte para notibuk	celulares		moda jovem	
15	teclado positivo		guarda roupa		blusa cropped	blusa croped
16	teclado		ps3		moda praia	
17	fonte lenovo g460		mini system	mini sistem	anitta	anita
18	fonte dell inspiron 1525		geladeira		tenis	
19	e1-531		samsung galaxy	samsumg galaxi	regata	
20	cabo flat	cabu flet	ventiladores		chinelo	xinelo

Quadro 2.3 – Termos utilizados para as coleta de dados

	Calçados	Variações fonéticas	Suplementos Alimentares	Variações fonéticas
1	usaflex		bcaa	
2	ramarim		whey	whei
3	bottero	botero	creatina	
4	crysalis	crisalis, crisalys	whey protein	whei protem
5	carmim		albumina	abumina
6	vizzano	vizano	lipo 6	
7	capodarte		glutamina	
8	itapuã		coqueteleira	

9	dakota		carnivor	
10	ferrette	ferete	dilatex	
11	via marte		nutrilatina	
12	tabita		colageno	
13	carrano	carano	max titanium	max tataniu
14	picadilly	picadily, picadili	whey gold	whei goldi
15	democrata		whey protein gold standard	whei goldi protem
16	luiza barcelos		probiotica	
17	moleca		caseina	
18	bebece		force 1	forci 1
19	kenner	kener	ultra whey gainer	ultra whei ganer
20	mizuno	misuno	zma	

Quadro 2.4 – Termos utilizados para as coleta de dados

Todas seguiram as regras a seguir:

a) Foi mantida uma média de vinte e cinco mil registros com uma margem de 5% para mais ou para menos, dependendo do ramo de atividade de cada loja, acessados através de um arquivo XML.
b) Os resultados referentes aos resultados das consultas foram considerados do seguinte modo:

Satisfatório	Termo pesquisado entre a 1ª e 12ª posição
Razoável	Termo pesquisado entre 13ª e 20ª posição
Insuficiente	Termo pesquisado entre 21ª e 40ª posição
Insatisfatório	Termo pesquisado acima da 41ª posição

c) Não foram consideradas variações de posição para classificação desde que o resultado apresentado para termos pesquisados com erros de digitação e erros de fonética se mantivessem na mesma faixa de classificação resultante originalmente da pesquisa pelo termo correto.

d) A faixa de corte para definição de um algoritmo assertivo foi estabelecida quando 80% dos resultados apresentaram o termo como satisfatório, cobrindo assim os 75% de abandono da loja na primeira pesquisa com resultados insatisfatórios.

Os testes foram feitos em ambiente de desenvolvimento e pesquisa para alcançar a melhor assertividade possível antes da implantação final do algoritmo de busca fonética nas lojas virtuais.

2.3.1 – APLICANDO O ALGORITMO SOUNDEX

Soundex original para o idioma Inglês

```
function Soundex ($palavraChave)
{
	for($i=0;$i -lt $palavraChave.Length;$i++)
	{
		if ($i -eq 0)
		{
			$soundexCode = [string]$palavraChave[0]
		}
		else
		{
			switch -regex ($palavraChave[$i])
			{
				"[bfpv]"	{ if ($soundexCode.Substring($soundexCode.Length - 1) -ne "1" ) { $soundexCode = $soundexCode + "1" } }
				"[cgjkqsxz]"	{ if ($soundexCode.Substring($soundexCode.Length - 1) -ne "2" ) { $soundexCode = $soundexCode + "2" } }
				"[dt]"		{ if ($soundexCode.Substring($soundexCode.Length - 1) -ne "3" ) { $soundexCode = $soundexCode + "3" } }
				"[l]"		{ if ($soundexCode.Substring($soundexCode.Length - 1) -ne "4" ) { $soundexCode = $soundexCode + "4" } }
				"[mn]"		{ if ($soundexCode.Substring($soundexCode.Length - 1) -ne "5" ) { $soundexCode = $soundexCode + "5" } }
				"[r]"		{ if ($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode = $soundexCode + "6" } }
			}
```

```
            }
        }

        $cutPoint = if ($soundexCode.Length -gt 4) { 4 } else {
$soundexCode.Length }
        return $soundexCode.Substring(0, $cutPoint)
}
```

Deste modo, é possível observar que os resultados obtidos com o Soundex se mostraram insatisfatórios. O nível de assertividade apresentado nos testes para o idioma brasileiro mostrou que 30% dos termos procurados eram encontrados na primeira pesquisa, porém em sua composição encontrava-se palavras originadas do idioma inglês.

Resultado final: insatisfatório.

2.3.2 – APLICANDO O ALGORITMO SOUNDEX MODIFICADO PARA O IDIOMA PORTUGUÊS.

Como o Soundex é um algoritmo voltado para o idioma inglês, acaba gerando muitos erros quando utilizado em outro idioma. Por isso, é necessário agrupar alguns conjuntos de letras por seus respectivos fonemas. O algoritmo BuscaBR foi o ponto de partida para essa modificação. Os demais agrupamentos fonéticos foram mantidos. Também foram aplicadas as regras do BuscaBR, como converter todas as letras para Maiúsculo, além de alguns agrupamentos de fonemas e eliminações de vogais, consoantes e terminações. O Quadro 3.0 demonstra essas substituições, eliminações e agrupamentos de fonemas.

Substituição	Eliminação
Y por I	Acentos
BR por B	Terminações S, Z, R, M, N, AO, L
GR, MG, NG, RG por G	Vogais
CA, CO, CU, por K	H
LH por L	Letras duplicadas
RM, GM, MD, SM por M	
NH por N	
PR por P	
TS, RS por S	

LT, TR, CT, RT, ST por T	
W por V	

Quadro 3.0 – Agrupamento e eliminações aplicadas na adaptação do algoritmo Soundex para atender o idioma português brasileiro.

```
function SoundexModificado ($palavra)
{

	$palavraChave = $palavra.toUpper();
	$palavraChave.Replace('Y','I');
	$palavraChave.Replace('BR','B');
	$palavraChave.Replace('GR','G');
	$palavraChave.Replace('MG','G');
	$palavraChave.Replace('NG','G');
	$palavraChave.Replace('RG','G');
	$palavraChave.Replace('CA','K');
	$palavraChave.Replace('CO','K');
	$palavraChave.Replace('CU','K');
	$palavraChave.Replace('LH','L');
	$palavraChave.Replace('RM','M');
	$palavraChave.Replace('GM','M');
	$palavraChave.Replace('MD','M');
	$palavraChave.Replace('SM','M');
	$palavraChave.Replace('NH','N');
	$palavraChave.Replace('PR','P');
	$palavraChave.Replace('TS','S');
	$palavraChave.Replace('RS','S');
	$palavraChave.Replace('LT','T');
	$palavraChave.Replace('TR','T');
	$palavraChave.Replace('CT','T');
	$palavraChave.Replace('RT','T');
	$palavraChave.Replace('ST','T');
	$palavraChave.Replace('W','V');

	for($i=0;$i -lt $palavraChave.Length;$i++)
	{
		if ($i -eq 0)
		{
			$soundexCode = [string]$palavraChave[0]
		}
		else
		{
			switch -regex ($palavraChave[$i])
			{
				"[BFPV]"          { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "1" ) { $soundexCode =
$soundexCode + "1" } }
				"[CÇGJKQSXZ]"     { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "2" ) { $soundexCode =
$soundexCode + "2" } }
				"[DT]"                  { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "3" ) { $soundexCode =
$soundexCode + "3" } }
				"[L]"             { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "4" ) { $soundexCode =
$soundexCode + "4" } }
```

```
                        "[MN]"                  { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "5" ) { $soundexCode =
$soundexCode + "5" } }
                        "[R]"               { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "6" } }

                    }
                }
        }

        $cutPoint = if ($soundexCode.Length -gt 15) { 15 } else {
$soundexCode.Length }
        return $soundexCode.Substring(0, $cutPoint)
}
```

Os resultados obtidos com as modificações apresentadas se mostraram em sua maioria insuficiente para as lojas virtuais do ramo de suprimentos de informática, no qual 71% dos resultados apresentados encontravam-se entre a 21ª e a 40ª posição, uma vez que os termos pesquisado são, em sua maioria, códigos. Isto é, eliminando os números que não se encontram na primeira casa de uma sequência, diminui-se a assertividade.

Já nas demais lojas virtuais analisadas, o nível de assertividade foi satisfatório e o algoritmo testado foi positivo para a tarefa. Os testes demonstraram 94% de acertos entre os doze primeiros resultados.

Resultado final: satisfatório / insuficiente.

2.3.3 – APLICANDO O ALGORITMOS SOUNDEX MODIFICADO PARA ANTEDER O PORTUGUÊS BRASILEIRO E CÓDIGOS NUMÉRICOS.

Soundex para atender o idioma português brasileiro e códigos numéricos.

```
function SoundexModificadoNumero ($palavra)
{

        $palavraChave = $palavra.toUpper();
        $palavraChave.Replace('Y','I');
        $palavraChave.Replace('BR','B');
        $palavraChave.Replace('GR','G');
        $palavraChave.Replace('MG','G');
        $palavraChave.Replace('NG','G');
        $palavraChave.Replace('RG','G');
        $palavraChave.Replace('CA','K');
        $palavraChave.Replace('CO','K');
        $palavraChave.Replace('CU','K');
        $palavraChave.Replace('LH','L');
        $palavraChave.Replace('RM','M');
```

```
        $palavraChave.Replace('GM','M');
        $palavraChave.Replace('MD','M');
        $palavraChave.Replace('SM','M');
        $palavraChave.Replace('NH','N');
        $palavraChave.Replace('PR','P');
        $palavraChave.Replace('TS','S');
        $palavraChave.Replace('RS','S');
        $palavraChave.Replace('LT','T');
        $palavraChave.Replace('TR','T');
        $palavraChave.Replace('CT','T');
        $palavraChave.Replace('RT','T');
        $palavraChave.Replace('ST','T');
        $palavraChave.Replace('W','V');

        for($i=0;$i -lt $palavraChave.Length;$i++)
        {
              if ($i -eq 0)
              {
                    $soundexCode = [string]$palavraChave[0]
              }
              else
              {
                    switch -regex ($palavraChave[$i])
                    {
                          "[BFPV]"          { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "1" ) { $soundexCode =
$soundexCode + "1" } }
                          "[CÇGJKQSXZ]"     { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "2" ) { $soundexCode =
$soundexCode + "2" } }
                          "[DT]"                  { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "3" ) { $soundexCode =
$soundexCode + "3" } }
                          "[L]"             { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "4" ) { $soundexCode =
$soundexCode + "4" } }
                          "[MN]"                  { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "5" ) { $soundexCode =
$soundexCode + "5" } }
                          "[R]"             { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "6" } }
                          "[0]"             { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "000" } }
                          "[1]"             { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "111" } }
                          "[2]"             { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "222" } }
                          "[3]"             { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "333" } }
                          "[4]"             { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "444" } }
                          "[5]"             { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "555" } }
```

```
                        "[6]"               { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "666" } }
                        "[7]"               { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "777" } }
                        "[8]"               { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "888" } }
                        "[9]"               { if
($soundexCode.Substring($soundexCode.Length - 1) -ne "6" ) { $soundexCode =
$soundexCode + "999" } }


                    }
                }
            }

            $cutPoint = if ($soundexCode.Length -gt 15) { 15 } else {
$soundexCode.Length }
            return $soundexCode.Substring(0, $cutPoint)
}
```

A partir dos resultados obtidos com a modificação do algoritmo Soundex para o português brasileiro, identificou-se outra deficiência: o problema de pesquisas com termos que contêm números em sua composição.

Para contornar este problema, foram considerados os números de uma sequência, repetindo cada casa decimal duas vezes e a faixa de corte do código orinalmente aplicada a quatro caracteres passou a ser de 15 caracteres para atender termos que contem uma letra seguida de até quatro numeros, o que evita conflitos com os códigos fonéticos. Assim, as Chávez para os termos como “C535” de C25 passaram a ser C555333555; com essa modificação a melhoria assertiva fica evidente quando codificamos o termo “calça 41” e “calça 4”, no algoritmo original o resultado para esse termo é C42 para ambos os termos; no algoritmo modificado para diferenciar numeros o resultado é C42444111 e C42444 respectivamente diferenciando os dois termos e se mostrando mais preciso.

Os resultados obtidos se mostraram satisfatórios: 96% dos termos pesquisados apresentaram seus resultados entre os doze primeiros itens encontrados, independente do ramo da atividade da loja virtual analisada. Ou seja, esta demonstrou ser a melhor solução encontrada para aplicação de um algoritmo fonético em uma loja virtual.

Resultado final: satisfatório.

2.3.4 – DIFICULDADES ENCONTRADAS

Em muitos casos os algoritmos fonéticos retornam a mesma chave para termos totalmente diferentes. Isso pode ser um problema principalmente em bancos de dados com muitos registros de itens cuja aplicação real é diferente, como “bola” e “bala”, “tecidos” e “tecidas”. Esses exemplos demonstram termos totalmente diferentes, mas para os quais as chaves fonéticas são exatamente as mesmas.

Não foi encontrada uma solução para modificar o algoritmo fonético de forma a resolver o problema de chaves iguais para termos diferentes.

3 – CONCLUSÃO

Com base nas análises realizadas conclui-se que é extremamente necessário adaptar os algoritmos fonéticos para o idioma ao qual se deseja aplicá-lo, bem como ao ramo de atividade da loja virtual. Como foi possível perceber na realização deste trabalho, as palavras podem ser erroneamente homófonas de um idioma para outro, e isso resulta em uma interpretação equivocada do algoritmo, que cria as chaves com base no som dos termos.

Além da adaptação ao idioma, também é preciso entender o comportamento do usuário dentro de uma loja virtual. É preciso saber e considerar quais são os termos mais pesquisados, a forma em que são pesquisados. As análises realizadas neste trabalhar permitem concluir que somente a fonética não atende o problema de sistemas que trabalham com todas as possibilidades de pesquisas, independente de código, nome ou fabricante.

4 – REFERÊNCIAS

1. IBGE, Censo 2011, http://www.ibge.gov.br, acessado em 03/09/2012.
2. Phonect Algorithm, http://en.wikipedia.org/wiki/Phonetic_algorithm, acessado em 09/10/2012.
3. SQL, http://wikipedia.org, acessado em 09/10/2012.
4. FBits One Stop Shop, Soluções F-Search Fbits, Novembro, 2011
5. Boosting Conversion Through Relevation Site Search, VerticalWeb Media, Unit SouthWacker, April 2008.
6. Soundex, http://www.avotaynu.com, acessado em 22/10/2012.
7. The Soundex Indexing System, The U.S. National Archives and Records Administration