Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

THE FUNCTION OF N-GRAMS SUBSTITUTION AND CODE SHIFT IN THE SOUNDEX ALGORITHM

Sri Nurdiati[1], Julio Adisantoso[1], Yeni Herdiyeni[1], R Zainal Arifin Fandi Saputra[1]
[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Bogor Agriculture Institute

ABSTRACT

The unclearness of word-root due to user limitation in the information of scientific names as well as the characteristic of scientific name retrieval system reduces the performance of the system. The objective of the research is to analyze the effect of n-grams substitution and code shift to increase the recall and precision value of the Soundex algorithm. For that intention, the following steps are conducted: make the scientific names dictionary, identify the scientific names in document, and rank the process by using a dice coefficient. The testing process uses 849 document collection and 20 kinds of query with different kinds of mistake. The performance of the retrieval is compared between using and not using n-grams substitution and a code shift, only using n-grams substitution (NS), and using both n-grams substitution and a code shift (CS).

The result of the research shows that the use of n-grams substitution and code shift is able to increase the performance of the scientific name retrieval system. Both techniques can retrieve up to 95% scientific names with 20 different queries.

The result of the research also shows that data would not affect the language when n-grams substitution and code shift are used. This is because n-grams substitution makes the change of the sound more uniform as a result of the match between two or more alphabets into one or more alphabets

Keyword: Information retrieval, scientific name, Fuzzy Soundex, Soundex, code shift, n-grams substitution, and dice coefficient.

1. INTRODUCTION

Name is an important thing in information system. It is often used as a search criteria in information retrieval system used in libraries (author), police files (prisoners), bookstores, etc.

Information retrieval system which produces a group of scientific name often discovers problems for queries in natural language context. The unclearness of word-root as a result of user limitation in knowing the information is a key problem of the system. Misspelling in writing the query will result in the failure of retrieving the information.

The misspelling problem can be solved by using phonetic similarity algorithms, such as Soundex, Phoenix, Pfeifer, and Fuzzy Soundex. Soundex algorithm has been developed for Indonesian query by modifying the group of consonants that appropriate with the Indonesian rule. A research conducted by Primasari in 1997 used a binary conversion for ranking system. The result showed that data did not effect the language. In this case, the Soundex algorithm which was developed for English is also also worked well with Indonesian. It is because both English and Indonesian classify their consonants using the similar extraordinary way.

Fuzzy Soundex algorithm has a better recall and precision value than other phonetic similarity algorithms. The improvement is caused by the two methods that inserted in Fuzzy Soundex : n-grams substitution and code shift. Both methods can increase the recall and precision value of phonetic similarity retrieval system.

Scientific name has two possibilities retrieval's result : can be retrieved or can not be retrieved at all. Scientific name has a unique characteristics as compared to other names. Scientific name has only one spesific name, while other names may have many similarities. The difference causes the sensitiveness of the scientific name retrieval system.

Based on the above explanation, the objective of the research is to analyze the effect of n-grams substitution and code shift to increase the recall and precision of the Soundex algorithm.

2. RELATED RESEARCH

Soundex Algorithm

Soundex is a phonetic algorithm that used to decrease the query mistyping effect of misspelling. The genuine Soundex algorithm has been patented by Margaret O'Dell and Robert C. Russell in 1918. The method that used is classifying 6 phonetic classification from human sound (bilabial, labiodental, dental, alveolar, velar, and glottal). The classification is based on lip and tongue position to make the sounds [4].

The Soundex algorithm is as follows:

1. Capitalize all letters in the word and drop all punctuation marks.
2. Retain the first letter of the word
3. Conversion process is based on Table 1.
4. Remove all pairs of digits.
5. Remove all zeros from the string.

ISSN 1858-1633 @2006 ICTS

© Hak cipta milik IPB, Institut Pertanian Bogor
Bogor Agricultural University

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

© Hak cipta milik IPB, Institut Pertanian Bogor
Bogor Agricultural University

6. Pad the string resulted from step (5) into only 4 bits, if it has less than 4 bits then adds with zeros, while if it has more than 4 bits then return only the first four positions. At the end it will be of the form <uppercase letter> <digit> <digit> <digit>.

Table 1 Consonant Classification of Some Soundex Algorithms

Alphabet	Soundex	Soundex Primasari (1997)	Fuzzy Soundex
F, V	1	1	1
B, P	1	6	1
S, Z	2	2	2
D, T	3	6	3
L	4	3	4
M, N	5	5	5
R	6	4	6
G, J, K, Q	2	6	7
X	2	2	7
C	2	6	2

In 1997, Soundex algorithm has been developed into Indonesian by re-modifying the consonant classification based on the following factors:
1. The articulator and point of articulation.
2. The way that passed by air.
3. The kinds of obstacle that found when the air was out.

In 2002, Dave Holmes and Catherine Mc Cabe modified Soundex algorithm by adding some methods into it. Then they name it Fuzzy Soundex. The methods can improve the recall and precision value of the retrieval system. Fuzzy Soundex blurs the query into some different codes. The phonetic retrieval system could increase the similarity measure between two names if the blur codes are in increasing number. In the case that each name only has one Soundex code, the similarity is binary, so that the similarity measure is not good enough (Holmes Dave March 9th 2006, private communication).

The code length and consonant classification in Fuzzy Soundex algorithm is different from the Soundex algorithm. The code length in Fuzzy Soundex is 5 bits, because by adding the code length by 1 bit, the mistake in the end of the name can be identified. Fuzzy Soundex uses n-grams substitution and code shift to increase the recall and precision value of the retrieval system.

Table 2 Mistake Cassification by Damerau

Type of Error	Baseline Name	Deviation
Insertion	Averrhoa	Averrkhoa
Omission	Retrofractum	Retrofactum
Substitution	Canna	Kanna
Transposition	Phyllanthus	Pyhllanthus

N-grams Substitution

Damerau defines misspellings into 4 categories as in Table 2 [2]. N-grams substitution can decrease the misspellings by substituting the characters that belong to substitution mistakes case. Table 3 illustrates types of n-grams that belong to substitution mistakes case [2].

Code Shift

Code shift is an effort to decrease insertion and omission type of mistakes. This method can identify the mistake at the beginning of the name, while for identifying the mistake at the end of the name is by adding 1 bit code length. Code shift can increase the recall value up to 96% by removing the second character from the five bits Fuzzy Soundex code [2].

Dice Coefficient

Dice coefficient is a formula to count the similarity measure between fuzzy codes. The formula is as follows [2]:

$$\delta = (2*\gamma)/(\alpha+\beta)$$

where :

δ is a dice coefficient

γ is common features

α is feature for name 1

β is feature for name 2

Table 3 Types of n-grams that Belong to n-grams Substitution Case

N-grams	Substitution
CA	KA
CC, CK, CH	KK
CE	SE
CL	KL
CR	KR
CI	SI
CO	KO
CS, CZ, TS, TZ	SS
CU	KU
CY	SY
DG	GG
GH	HH
GN, KN, NG	NN
HR, WR	RR
HW	WW
PF, PH	FF
SCH	SSS
TIO	SIO

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.
© Hak cipta milik IPB (Institut Pertanian Bogor)
Bogor Agricultural University

3. METHODOLOGY

Scientific name retrieval system model

The scientific name retrieval system model used in this research is shown in Figure1.

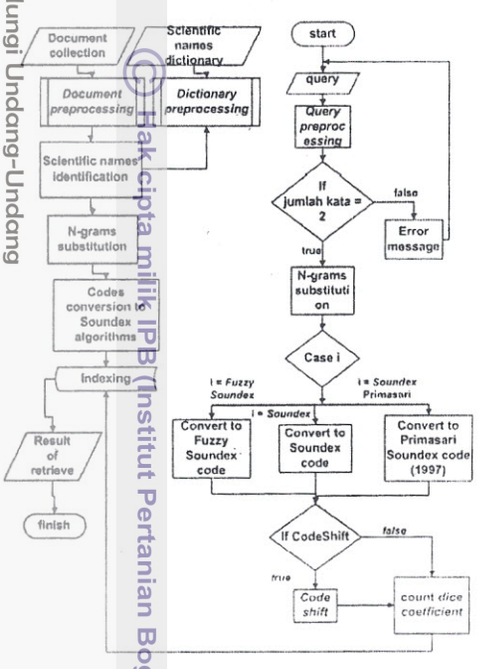

Figure 1 Scientific Name Retrieval System Model.

Dictionary preprocessing

Scientific names in the dictionary have different word length, so they needs to be equalized to make the searching process easier because the system will start searching from the first word.

Word length will be equalized into two words. The justifications are as follows:

1. Words after the second word in the scientific name are rarely used. Usually, these are just abbreviation from descriptors who theirs research are well-known [1]. For example: *Artemisia vulgaris Linn.*
2. Two words are based on a binomial nomenclature system For example: *Averhoa bilimbi* (Sour carambola). *Oryza sativa* (Rice plant).
3. Two words have guaranteed the uniqueness of 354 scientific names in the dictionary.

Fuzzy Code

Query that has been converted into Fuzzy Soundex code will be blurred by decreasing the code length in each iteration. For example, if the query is 'A2546', then the fuzzy codes are as follows (Holmes Dave March 16th 2006, private communication):

- A2546 (*fuzzy* 5)
- A254 (*fuzzy* 4)
- A25 (*fuzzy* 3)
- A2 (*fuzzy* 2)
- A546 (*code shifted*)

Treatments

Three kinds of Soundex algorithm that will be compared will be given the same treatment, which are:

- Using the same blurring process
- Setting the code length to only 5 bits
- Adding the substitution n-grams and code shift into the Soundex Algorithm
- Using dice coefficient for ranking algorithm.

Assumptions

- Scientific name in the document is always right, so that scientific name identify process will be passed.
- The relevant documents are documents that have scientific names.

4. RESULTS AND DISCUSSION

Scientific Names Identification Analysis

The identification process will be done by comparing the first word of the scientific names in the dictionary with words result from document preprocessing. If a word is found, then its position will be saved to check the second word based on that position. This process will be repeated until the word is found. There are some cases where some scientific names have similar front names, such as Averhoa carambola and Averhoa bilimbi. Both are carambola family but with different species.

The result of the identification process of 849 documents is 646 names. It takes about 7 minutes and 11 seconds to execute the process.

N-Grams Substitution

The difference between n-grams substitution and n-grams is on the text condition that will be corrected. N-grams will correct the text that has technical mistake, for example mistyping. N-grams substitution will correct the alphabets that the sound has change if it runs into another alphabets.

N-grams substitution has a great effect when mistyping happens in the beginning of typing. The first word which is not converted into a code, results in a name that has identical spelling in the beginning of the name will be considered as different codes.

Koleus skotiolariades is one of the misspelling names that has Coleus scutellarioides as its originall name. Character 'C' is often spelt as 'K' if it runs into vocals 'A', 'U', and 'O'. Koleus and Coleus codes are considered different although they have passed the fuzzy process (Koleus's Fuzzy code is 'K4200' and Coleus's Fuzzy code is 'C4200'). It is because the fuzzy process does not crop the character in the beginning of the word.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

© Hak cipta milik IPB (Institut Pertanian Bogor)

N-grams substitution substitutes the character based on substitution cases on Table 3. Before going through the code converter process, character 'C' in 'Coleus' will be substituted by character 'K' because character 'C' is followed by 'O'. Now, 'Coleus' has the same code as 'Koleus'.

Some n-grams substitution cases other than those of Table 3 are found in this research. User oftenly makes mistake if he spells characters, like 'NJ' and 'Z' followed by vocal. Character 'NJ' on Ficus Benjamina is often heard as Ficus Benyamina. The substitution n-grams addition case is on Table 4.

Table 4 The additional cases of n-grams substitution

N-grams	Substitution
NJ	NY
Z + 'A, I, U, E, O'	J + 'A, I, U, E, O'

Dice Coefficient

The result from blurring the queries and blurring words in document will be compared to find their intersection. The algorithm of the matching process is as follows:

For all first word of scientific names from document

Count the dice coefficient value, using diceCoef (token1, token2).

If the dice coefficient $\neq 0$, then save the document id and the dice coefficient value.

Below is the dice coefficient algorithm,

diceCoef (token1, token2)

1. Blur token1 and token2.
2. Initialize the common value by 0.
3. For all the first word fuzzy queries
4. For all the first word fuzzy tokens.
5. If fuzzyQuery1 = fuzzyToken1, then the common value is added by 1.
6. For all the second word fuzzy queries.
7. For all the second word fuzzy tokens.
8. If fuzzyQuery2 = fuzzyToken2, then the common value is added by 1.
9. Count the dice coefficient using:

   dCoef=(2*common)/((fQuery1.length*2)+(fToken1.length*2))
10. Return dCoef value.

If there is a significant mistake in the first word, while in the second word is an unsignificant mistake, the algorithm is still able to retrieve the information. For example Kromotoli penata (the original name is Quamoclit penata). In this case the code from Kromotoli and Quamoclit is different.

Table 5 The Comparison of Kromotoli's and Quamoclit's codes

Kromotoli	*Quamoclit*
K6534	K5243
K653	K524
K65	K52
K6	K5
K534	K243

From Table 5 it can be seen that the two codes do not have similarity although they have been added by code shift, while the penata's code (the original name is pennata) has a close similarity (in Table 6).

Common value is a union of the intersection between the first and the second word. Although the common value of the first word is 0, common value of the second word can be very high, which is 5. The dice coefficient value can be determined as follows:

dCoef = (2*(0+5)) / ((5*2)+(5*2))
= 10 / 20
= 0.5

Table 6 The Comparison of Penata's and Pennata's codes

penata	*pennata*
P5300	P5300
P530	P530
P53	P53
P5	P5
P300	P300

The algorithm is still saving the document id and dice coefficient value because the value is more than 0, so that the document that related to pennata is still retrieved.

Code Shift

Code shift algorithm has a big effect on phonetic similarity retrieval system for insertion and omission type of mistakes. For example, asproha brimbi (the original name is Averrhoa bilimbi), has a 0 recall and precision value when it does not use a code shift. There are around 21 names which can be retrieved. A very different result was shown different when the code shift is used, where all the related documents were successfully retrieved.

Bogor Agricultural University

Table 7 The Effect of the Code Shift in Asproha brimbi

	Asproha brimbi		*Averrhoa bilimbi*	
Fuzzy 5	A2160	B6510	A1600	B4510
Fuzzy 4	A216	B651	**A160**	B451
Fuzzy 3	A21	B65	A16	B45
Fuzzy 2	A2	B6	A1	B4
Code shift	**A160**	**B510**	A600	**B510**

Table 7 shows that asproha brimbi was succesfully retrieved because the Apsroha's code, when code shift was added, is similar to that of Averrhoa's code in Fuzzy 4. The same case happens in brimbi's and bilimbi's code when the code shift was added.

Evaluation

To know the effect of n-grams substitution and code shift in increasing recall and precision value of Soundex algorithm, the result of the retrieval process was compared between use and not use n-grams substitution and code shift, only used n-grams substitution (NS), and used both substitution n-grams and code shift (CS).

The recall-precision curve in Figure 2 shows that n-grams substitution and code shift could increase the three of Soundex algorithm performance. The average of the increasing recall and precision value from three of Soundex algorithm is 20%.

The increasing of recall and precision value from three of Soundex algorithm is affected by the capability of n-grams substitution and code shift in handling various type of mistakes, such as insertion, omission, and transposition. The used of the two methods in the Soundex algorithm could increase the recall and precision value (for insertion and omission type of mistakes) in the average around 30% (Figure 3).

For transposition mistake, the average of the increase could reach around 40% (Figure 4). The use of the two methods in the Soundex algorithm which is investigated in the Primasari research (1997) could give a better result for this type of mistake. It is shown in the figure that the recall-precision curve of this Soundex algorithm almost reach a maximum value.

Figure 2 also shows that data does not give any effect to language when n-grams substitution and code shift are used because the maximum value that could be reached by the three Soundex curves were very close to each others. This is because n-grams substitution makes the change of the sound more uniform as a result of the match between two or more alphabets into one or more alphabets.

For example, 'C' tends to change into 'K' if runs into vocals 'A', 'U', and 'O', and change into 'S' if runs into vocals 'I' and 'E'. In Fuzzy Soundex algorithm, 'C' and 'K' are in different group, while the rest are in one group. As a comparison, in Soundex algorithm investigated by Primasari (1997), 'C' and 'S' are in different group, while the rest are one group.

The performance of the Soundex algorithm investigated by Primasari (1997) has increased in this research. The reasons can be explained as follows:

- The blurring process can increase the amount of the possibility of the related names to be retrieved.
- The used of the n-grams substitution and code shift and the increment of 1 bit code length.
- The used of the dice coefficient for ranking algorithm. In Primasari research (1997), the ranking algorithm used is a Binary Conversion.

Hak cipta dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.
© Hak cipta milik IPB (Institut Pertanian Bogor)
Bogor Agricultural University

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

© Hak cipta milik IPB (Institut Pertanian Bogor)

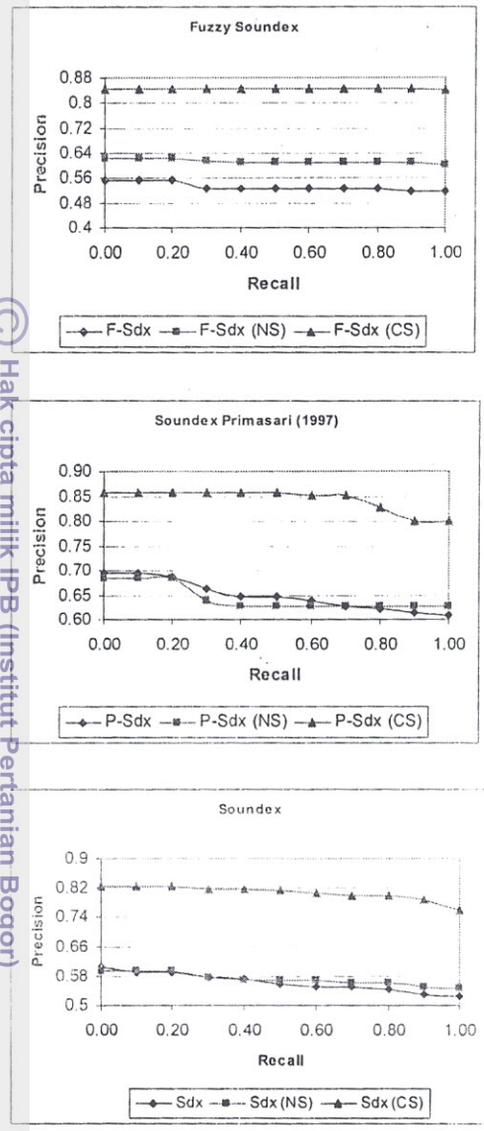

Figure 2 Recall-precision Curves of Some Soundex Algorithms

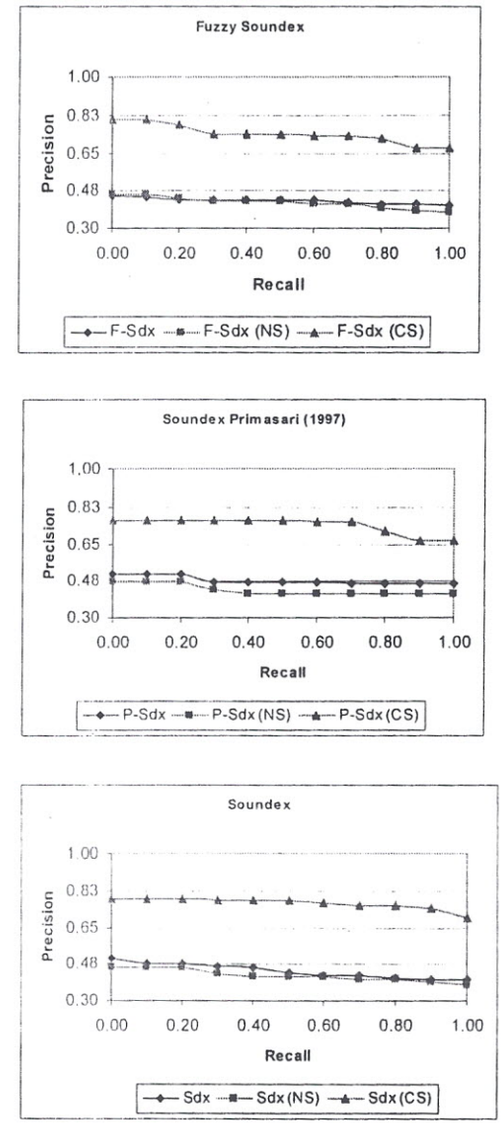

Figure 3 Recall-precision Curves of Some Soundex Algorithms (insertion and omission).

The positive effect of the n-grams substitution and code shift in Soundex algorithm can be used by Bogor Agriculture Institute's library in fixing its search system for scientific name queries. The mistakes occurred in typing scientific names, may cause the problem in retrieving the relevant information.

The condition of the searching system can disturb college students in searching information, since there are only a liminted number of college students who really understand scientific names form some species. One of the reasons is that it is quite difficult in spelling scientific names. College students who have just understood scientific names from some species, will still very likely make many mistakes in spelling or typing.

Bogor Agricultural University

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.
© Hak cipta milik IPB (Institut Pertanian Bogor)
Bogor Agricultural University

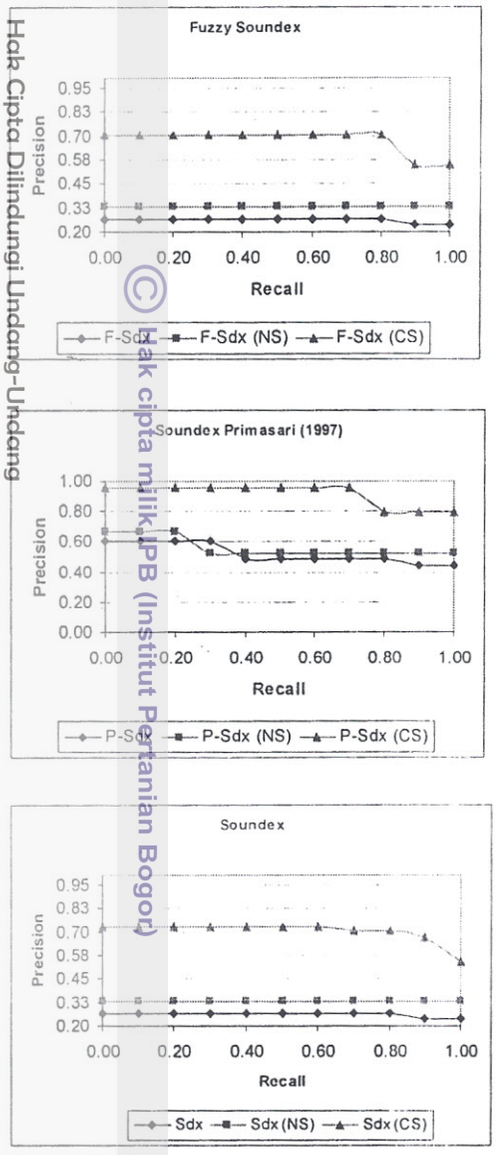

Figure 4 The Recall-precision Curves of Some Soundex Algorithms (transposition).

5. CONCLUSIONS

From the research it can be concluded that the insertion of n-grams substitution and code shift into a Soundex algorithm increases the recall and precision value of the scientific names retrieval system. From the testing queries, the insertion of both methods can retrieve up to 95% scientific names which have different type of mistakes.

In addition, it can be concluded that data would not affect the language when n-grams substitution and code shift are used. This is because n-grams substitution makes the change of the sound more uniform as a result of the match between two or more alphabets into one or more alphabets.

6. REFFERENCES

[1] Hendrawan Kiki. 2004. Cara Klasifikasi dan Tata Nama. http://clearinghouse.dikmenum.go.id/showContent.php?id=192&idCont=Bpn&SubjectID=21&mnMode=mnBp [21 Maret 2006].

[2] Holmes David, Catherine McCabe M. 2002. *Improving Precision and Recall for Soundex Retrieval.* Las Vegas. http://ir.iit.edu/publications/downloads/IEEESoundexV5.pdf [12 Juni 2005].

[3] Primasari Dewi. 1997. *Metode Pencarian dan Temu-Kembali Nama Berdasarkan Kesamaan Fonetik.* [Skripsi]. Bogor: Departement of Computer Sciences Fakultas Matematika dan Ilmu Pengetahuan Alam Institut Pertanian Bogor.

[4] Repici Dominic J. 2006. Soundex Algorithms Explained. http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm#Algorithm [14 Juni 2005].