Figure 1

Cleansing & Parsing & Standardization
Identify key components of inquiry data
Name, Address & City Normalization
Treat common abbreviations
Recognize diminutives
Treat plurals and other variants
Address Standardization
Vanity cities and streets
Known name changes in key regions
Candidate Retrieval
Create an optimized approach to retrieving best possible list of candidates
Tailor approach based on requirements for quality & speed
Best possible match candidates gathered from reference database
Evaluation & Decisioning
Evaluate matches according to consistent standards
MatchGrade pattern establishes granular feedback
Confidence Code enables auto-decisioning
Match Data Profile for extra information

Figure 2

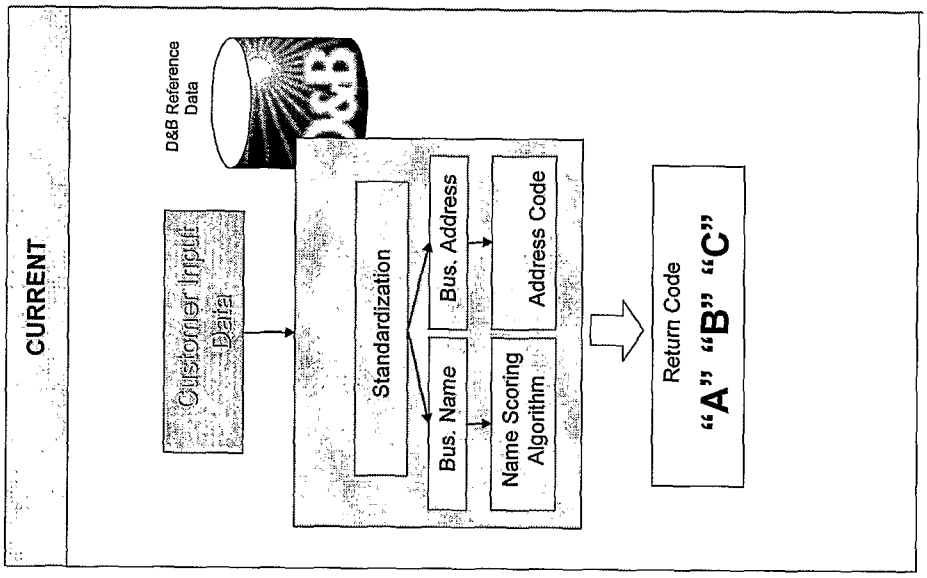

CURRENT
D&B Reference Data
Customer Input Data
Standardization
Bus. Name
Bus. Address
Name Scoring Algorithm
Address Code
Return Code
"A" "B" "C"

Figure 3

CC	Byline
10	"Very good matches"
7 - 9	"Good but not perfect" (i.e. due to missing inquiry data or variations in business *nomenclature*)
5- 6	"Worth considering" (i.e. there may be some opportunity mixed in with otherwise un-useful matches)
4 and below	"Probably not" (i.e. mostly un-useful, but some residual value)

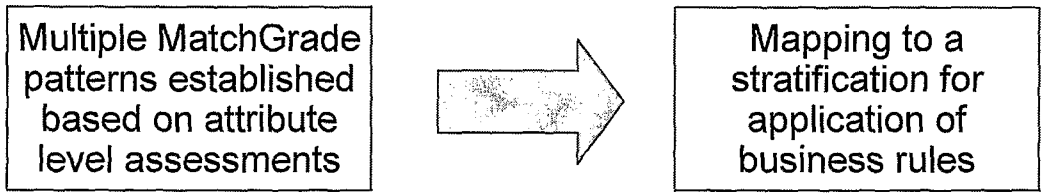


Figure 4

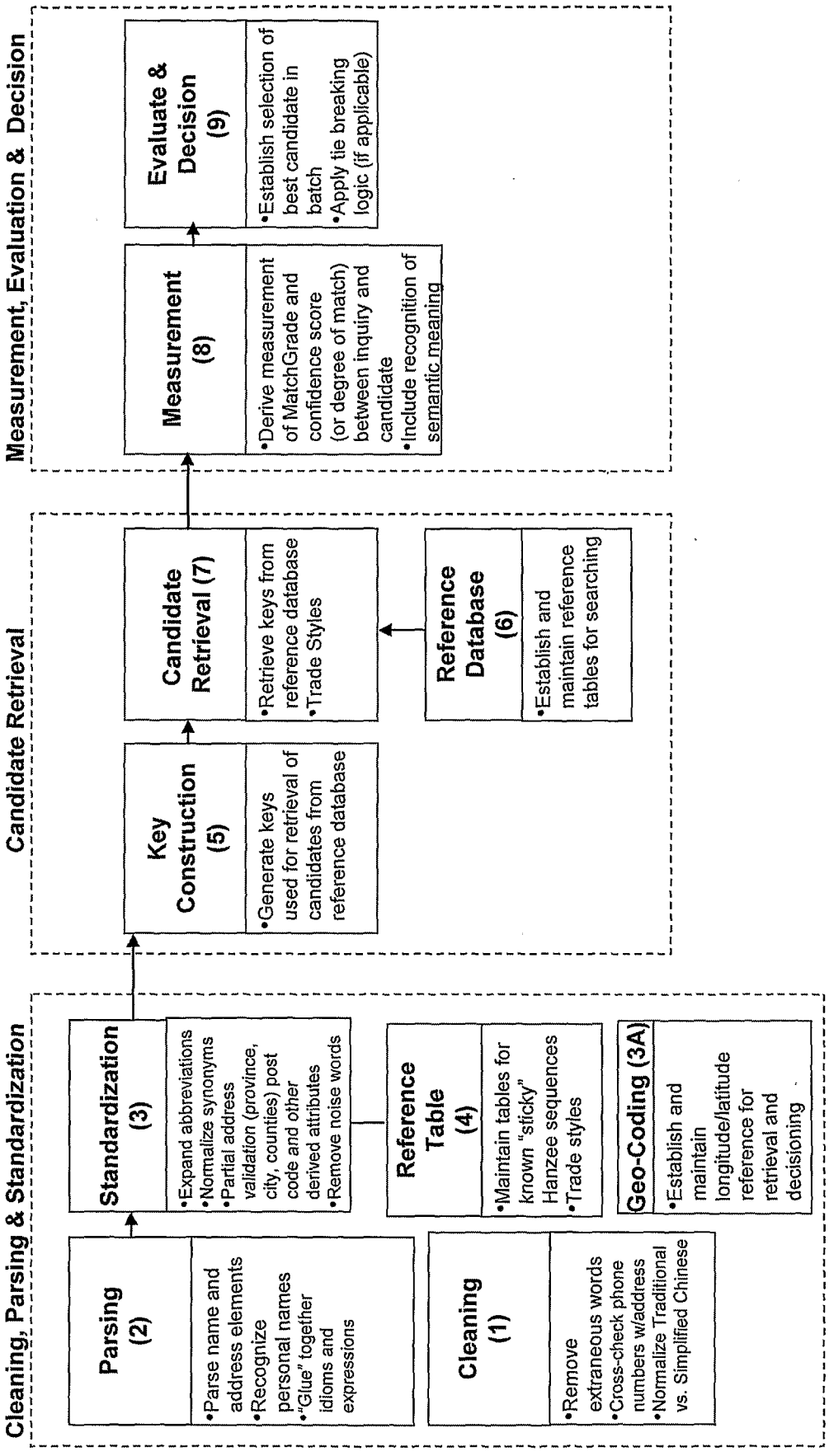

Cleaning, Parsing & Standardization
Parsing (2)
•Parse name and address elements
•Recognize personal names
•"Glue" together idioms and expressions
Standardization (3)
•Expand abbreviations
•Normalize synonyms
•Partial address validation (province, city, counties) post code and other derived attributes
•Remove noise words
Cleaning (1)
•Remove extraneous words
•Cross-check phone numbers w/address
•Normalize Traditional vs. Simplified Chinese
Reference Table (4)
•Maintain tables for known "sticky" Hanzee sequences
•Trade styles
Geo-Coding (3A)
•Establish and maintain longitude/latitude reference for retrieval and decisioning
Candidate Retrieval
Key Construction (5)
•Generate keys used for retrieval of candidates from reference database
Candidate Retrieval (7)
•Retrieve keys from reference database
•Trade Styles
Reference Database (6)
•Establish and maintain reference tables for searching
Measurement, Evaluation & Decision
Measurement (8)
•Derive measurement of MatchGrade and confidence score (or degree of match) between inquiry and candidate
•Include recognition of semantic meaning
Evaluate & Decision (9)
•Establish selection of best candidate in batch
•Apply tie breaking logic (if applicable)

Figure 5

CORRELATING JAPAN ADDRESS COMPONENTS TO PRIOR MATCHGRADE STRUCTURE

Field Name	Bus Name	Street #	Street Name	City	Province/Prefecture	PO Box	Tel	Postal Code	Density	Unique	Industry
Possible Values	A,B,F	A,B,F,Z	A,B,F,Z	A,B,F,Z	A,B,F,Z	TBD	A,B,F,Z	A,B,F,Z	A,B,F,Z	A,B,F,Z	A,B,F,Z
# values	3	4	4	4	4	1	4	4	4	4	4

Figure 6

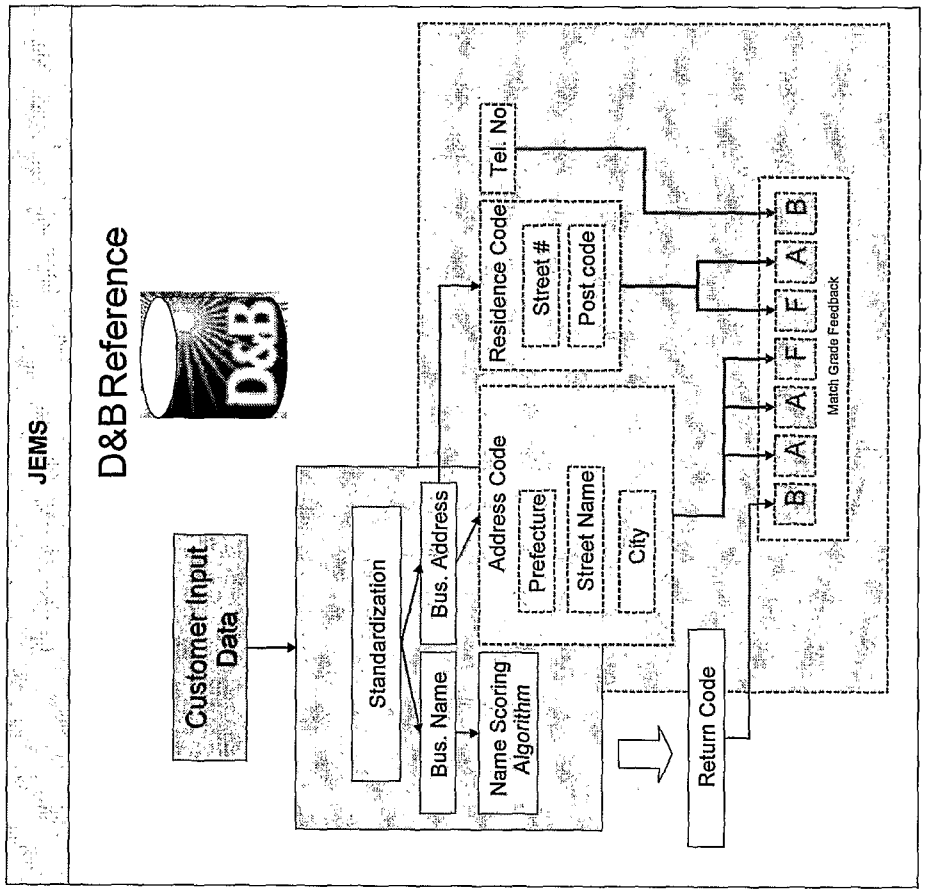
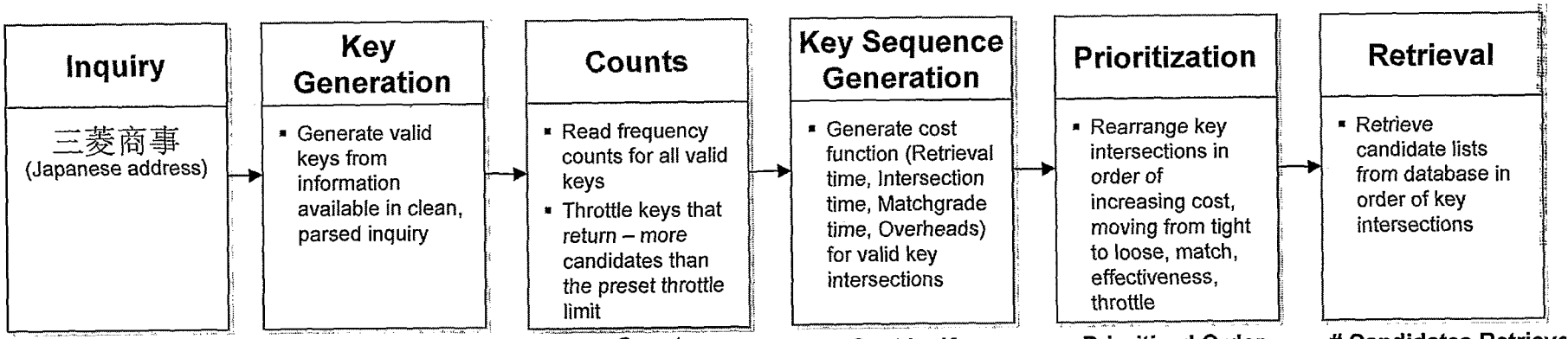

JEMS
D&B Reference
Customer Input Data
Standardization
Bus. Name
Bus. Address
Name Scoring Algorithm
Address Code
Prefecture
Street Name
City
Residence Code
Street #
Post code
Tel. No
Return Code
B A A F F A B
Match Grade Feedback

Figure 7

Matchgrade_Pattern	Confidence_Code	MGB01_Name	MGB02_StreetNum	MGB03_StreetName	MGB04_City	MGB04_Provlnce/Prefecture	MGB06_MailAddr	MGB07_Tel	MGB08_Postal Code	MGB09_Density	MGB10_Uniqueness	MGB11_Industry	Version	Update_date
AAAAAZAAZZZ	9	A	A	A	A	A	Z	A	A	Z	Z	Z	1.0	21/01/2005 09:28
AAAAAZABZZZ	7	A	A	A	A	A	Z	A	B	Z	Z	Z	0.5	1/01/2005 13:02
AAAAAZAFZZZ	7	A	A	A	A	A	Z	A	F	Z	Z	Z	0.5	1/01/2005 13:02
AAAAAZAZZZZ	7	A	A	A	A	A	Z	A	Z	Z	Z	Z	0.5	1/01/2005 13:02
AAAAAZBAZZZ	7	A	A	A	A	A	Z	B	A	Z	Z	Z	0.5	1/01/2005 13:02

Figure 8



Example	Key	Value	Count	Cost by Key	Prioritized Order	# Candidates Retrieve
	Exact Name	三菱商事	1	Exact Name 6.05	Exact Name	1
	Principle words	三菱 商事	3 2957	Principle Words 6.15	Principle Words	3
	Method 2	...	5	Method 2 809	Method 2	5
	Kingling	See figure 12 for derivation of "interesting dumb Kingles"	21 2,370 5,802 6,062 6,277 7,325 8,166	Kingle 665	Kingling	21
	Kacronym	...	590	Kacronym 102.15	Full Text Edit	1923
	Full Text index	See figure 23	1,923	Full Text edit 16.2	Kacronym	590

Example: (Cost by key is representative, algorithms are representative of the full range of available retrieval methods)

Figure 9

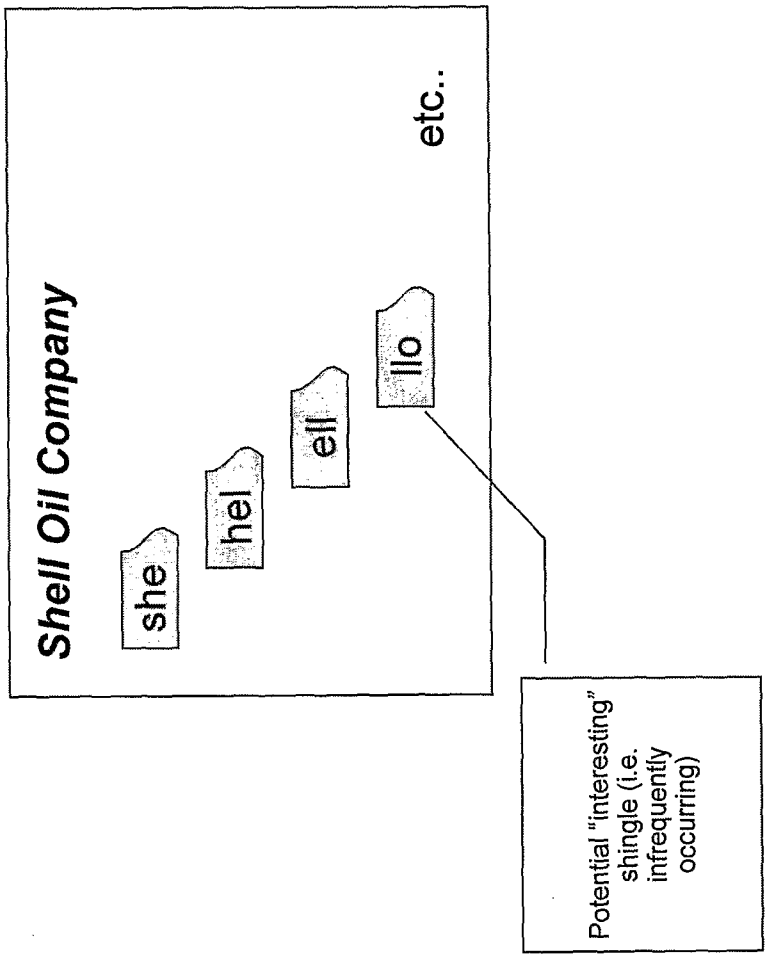
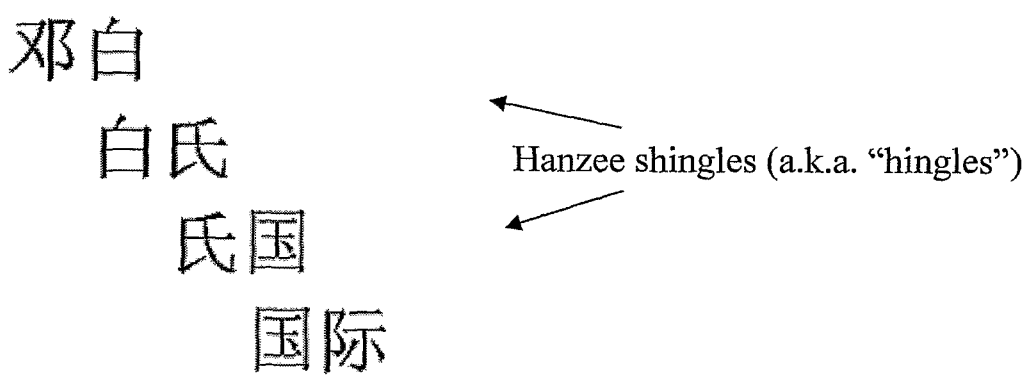

Shell Oil Company
she
hel
ell
llo
etc..
Potential "interesting" shingle (i.e. infrequently occurring)

Figure 10

HINGLES

邓白氏国际信息咨询（上海）有限公司



Note: Hingles are shown with a size of two and an offset of one for illustration purposes. Other hingle widths and offsets have different implications for semantic retention, predictiveness as a retrieval method, and computational complexity to derive.

Figure 11

STICKY HINGLES

邓白氏国际信息咨询（上海）有限公司

上 "up, above" + 海 "sea" = 上海 "Shanghai"

Allowing certain Hanzee characters to "stick" accounts for the need to retain meaning in the semantic aggregations.

Figure 12

DUMB HINGLES

Base String:

邓白氏国际信息咨询（上海）有限公司

Broken into “words”:

邓白氏 国际 信息咨询 上海 有限公司

信息 咨询 上海

all “Noisy” = low value

But if you look at Hingles that join “noisy” words.....

询上	▪ Much less common ▪ Very predictive!

Figure 13

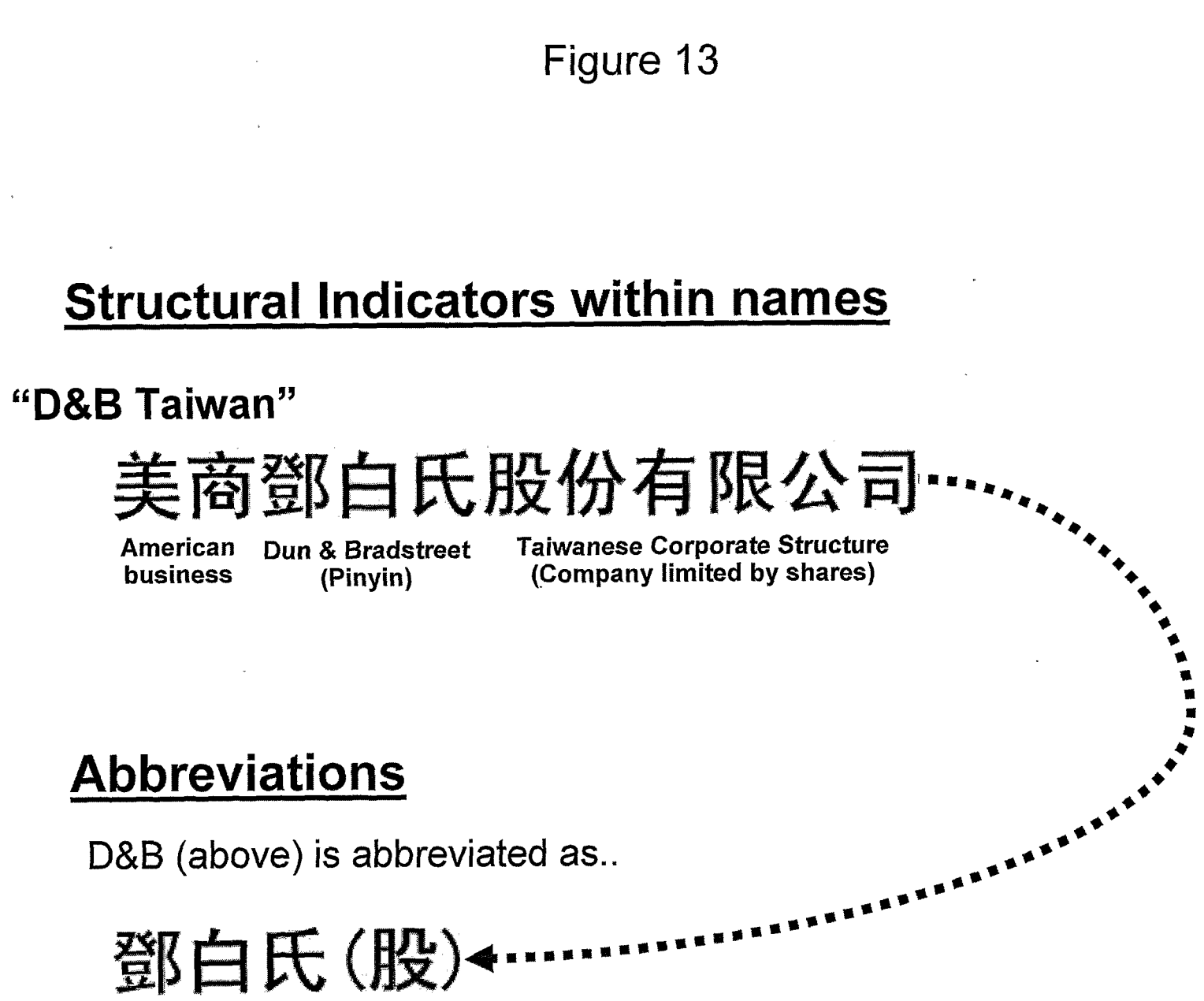

Structural Indicators within names
"D&B Taiwan"
美商鄧白氏股份有限公司
American business
Dun & Bradstreet (Pinyin)
Taiwanese Corporate Structure (Company limited by shares)
Abbreviations
D&B (above) is abbreviated as..
鄧白氏(股)

Figure 14

Divisions / Departments within the name.

宏基电子 Acer Electronics

宏基电机 Acer Machinery

宏基资讯 Acer Information Systems

Figure 15

HACRONYMS

"Hacronyms". (Hanzee Acronyms)

联华电子 / 联电 "United Electronics" "U.E."

Figure 16

TREATING SIMILARITIES ACROSS TRADITIONAL AND SIMPLIFIED SCRIPT

Traditional & Simplified Script

Generally:

Traditional	Simplified
Taiwan Hong Kong	China

But it is not always so clear-cut...

English:	D&B International Consultant	(Shanghai)	Co., Ltd
100% Simplified:	邓白氏国际信息咨询 Simplified	（上海） Same in Both	有限公司 Simplified
Mixed:	鄧白氏國際信息咨詢 Traditional	（上海）	有限公司 *Simplified*

Figure 17

EXAMPLES OF COMPUTING ALGORITHMIC CONSTANTS FOR RETRIEVAL OPTIMIZATION

A(n) Algorithm	CC Computational Complexity	P Predictiveness	AC Algorithmic Constant
A1	4	1/4	1.00
A2	9	1/3	3.00
A3	2	1/12	0.17

A(n) Algorithm	AC Algorithmic Constant	Y Expected Yield	X Score
A1	1.00	800	800.00
A2	3.00	50	150.00
A3	0.17	100	16.67

Figure 18

EVALUATION AND DECISIONING

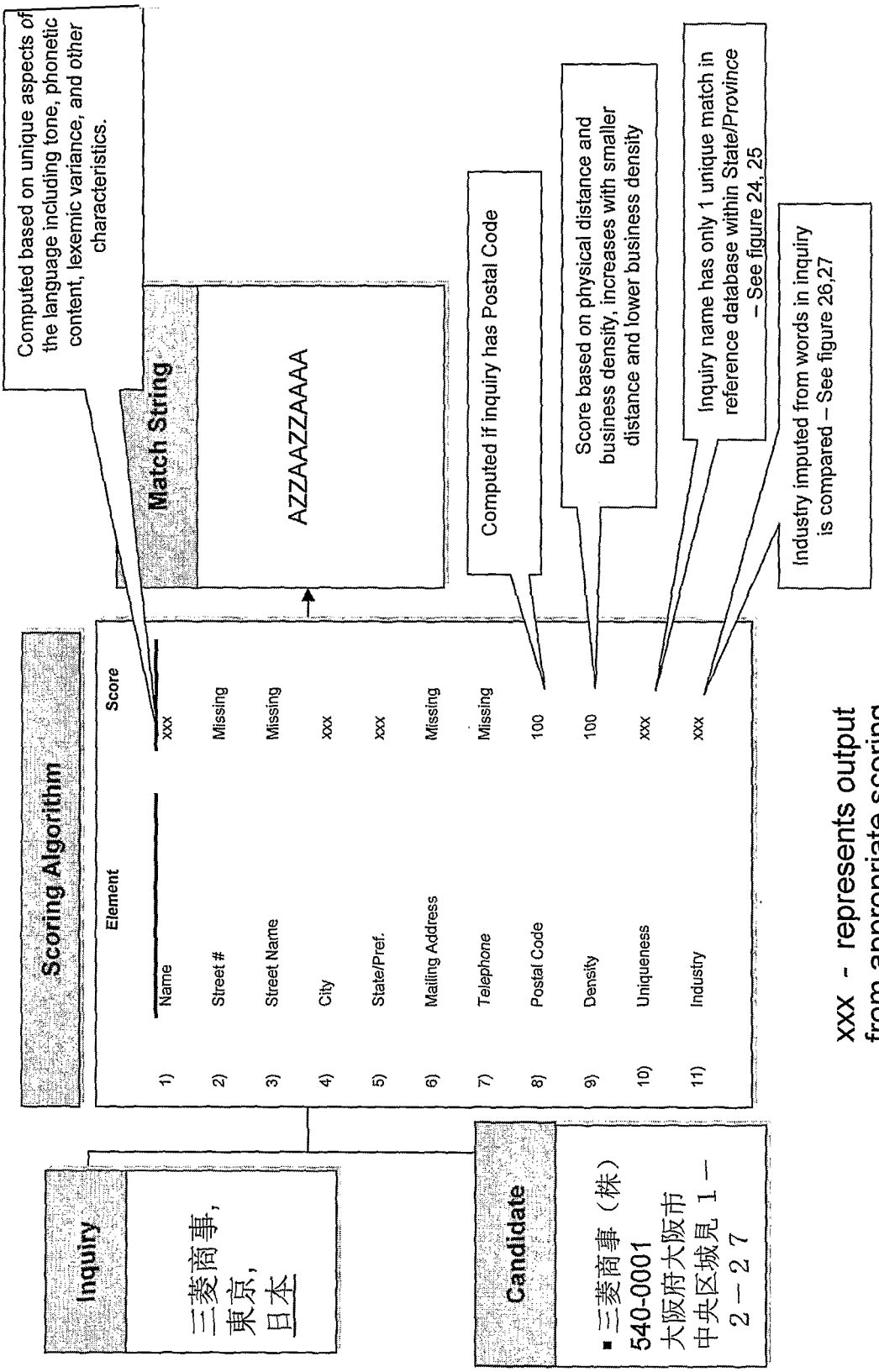


Figure 19

MATCH QUALITY FEEDBACK ATTRIBUTE – LEVEL EVALUATION

Business Name	Street Name	Province/ Prefecture	Telephone	*Density*	*Industry*
A	**A**	**A**	**Z**	*A*	*Z*
A	**A**	**B**	**B**	*F*	*B*

Street # City Secondary Address *Postal Code* *Uniqueness*

Figure 20

EXAMPLES OF ATTRIBUTE LEVEL EVALUATIONS

A	Same	三菱商事（株） vs.三菱商事	Mitsubishi Trading Corp. vs. Mitsubishi Trading
B	Similar	三菱商事（株） vs.三菱自動車工業（株）	Mitsubishi Trading Corp. vs. Mitsubishi Motors Corp
F	Different	三菱商事（株） vs.三井物産（株）	Mitsubishi Trading Corp. vs. Mitsui Corp
Z	Null		One or Both are Null (Blank)

Figure 21

STRATIFICATION OF ATTRIBUTE – LEVEL FEEDBACK

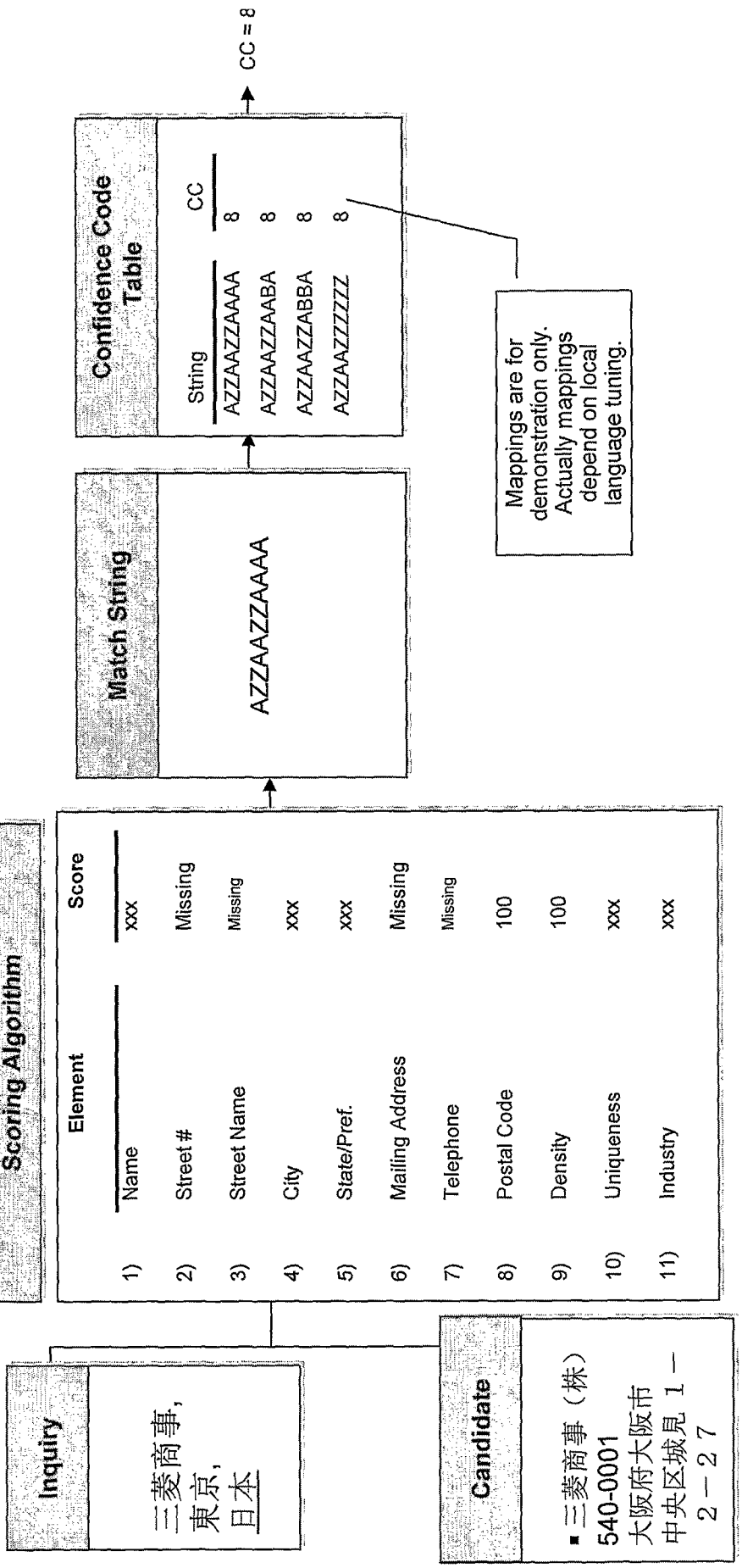


Figure 22

Sample non-editible report with Chinese data and English substrate

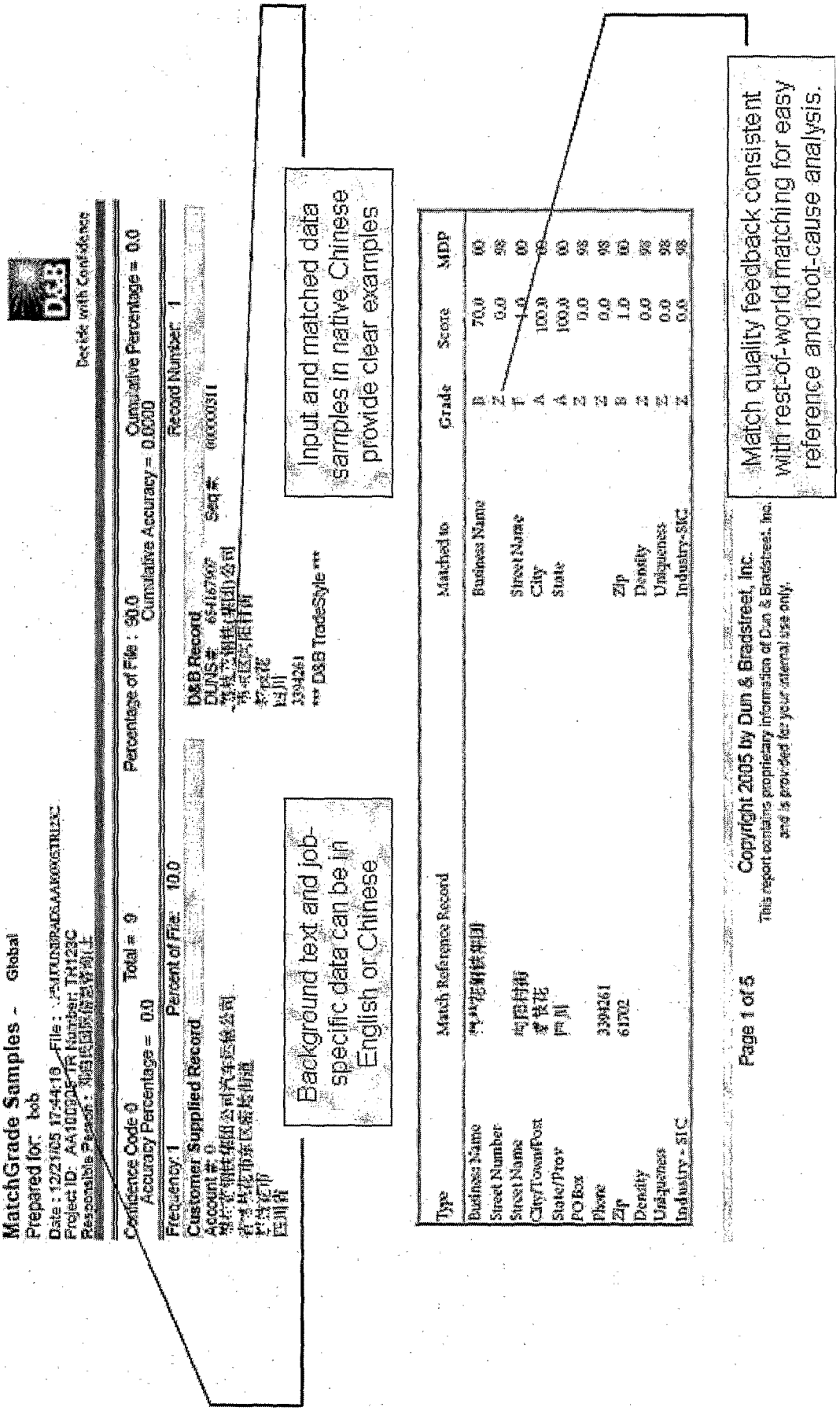


Figure 23

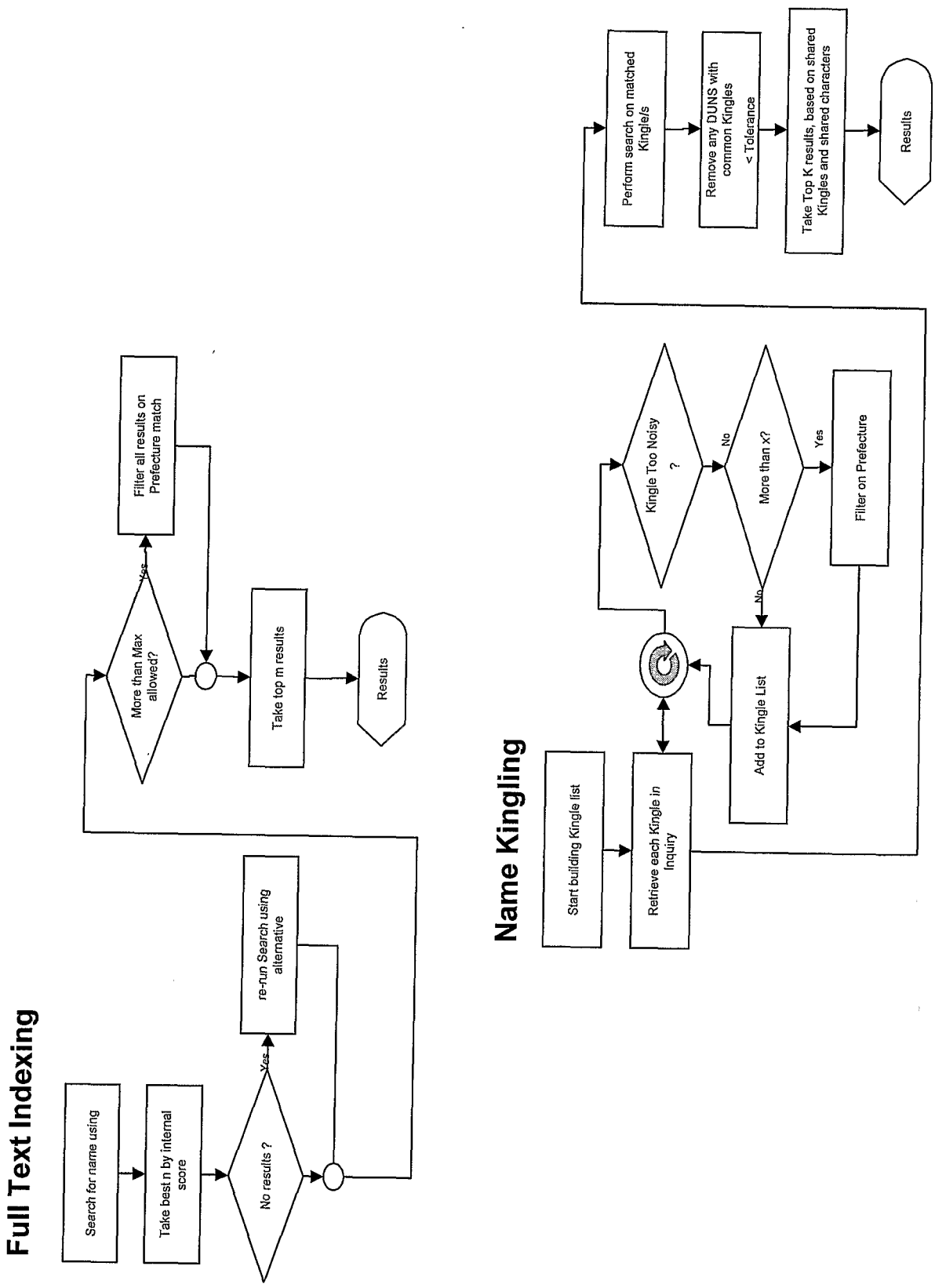
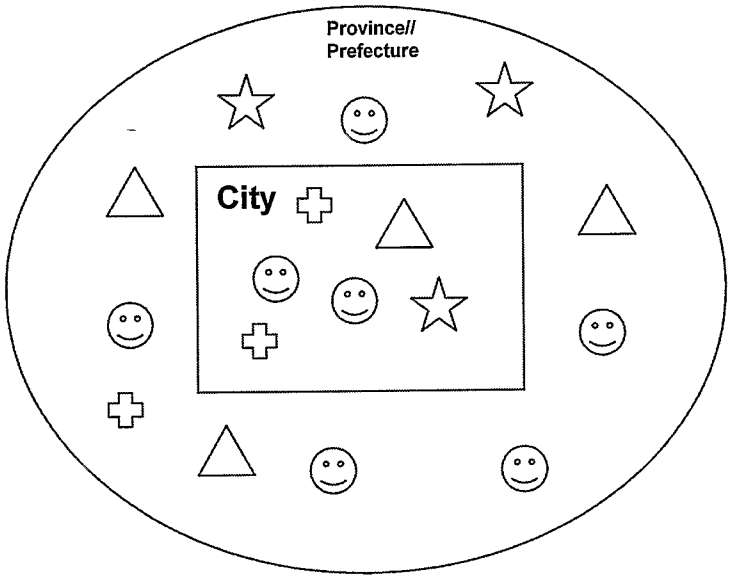

Full Text Indexing
Search for name using
Take best n by internal score
No results ?
Yes
re-run Search using alternative
More than Max allowed?
Yes
Filter all results on Prefecture match
Take top m results
Results
Name Kingling
Start building Kingle list
Retrieve each Kingle in Inquiry
Kingle Too Noisy ?
No
More than x?
Yes
Filter on Prefecture
No
Add to Kingle List
Perform search on matched Kingle/s
Remove any DUNS with common Kingles < Tolerance
Take Top K results, based on shared Kingles and shared characters
Results

Figure 24

Sample MatchGrade Algorithm UNIQUENESS

Uniqueness tells us how unique a company name is <u>within its geographical context</u>.



We score a candidate's uniqueness compared to an inquiry, based on a match between City and/or Province/Prefecture (or whatever geo units we use).

We consciously select <u>the smallest available geo common to both inquiry and candidate</u>.

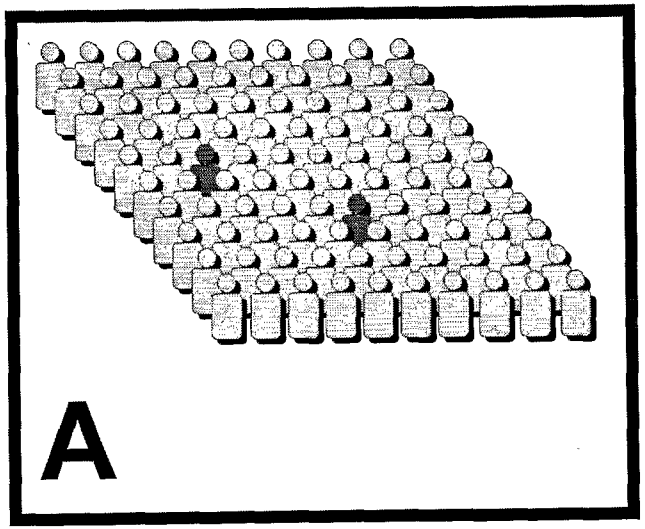
Size of Geo	# Businesses	Value of Being Unique
Small Geo	Few	Low
Large Geo	Many	High

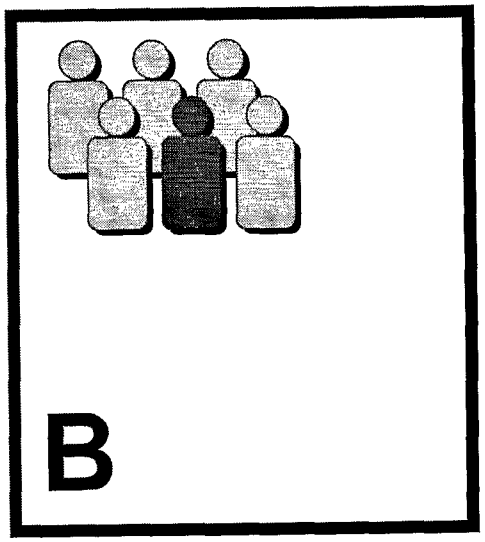
Assumptions:

1) The smaller area is always a subset of the larger one (eg city/state city/prefecture, province/country)
2) The size of the area is the count of businesses in that area (ie the number of entities in our reference file for that area)
3) Uniqueness is calculated according to an algorithm which is tuned to provide results mappable to A/B/F/Z

Figure 25

RELATIVE VS. ABSOLUTE UNIQUENESS





Size of Geo	# Businesses	Value of Being Unique
Small Geo	Few	Low
Large Geo	Many	High

According to NextGen the red person in **B** is more unique than the two in **A**, even though the total size of the B Geo is only 6.

But why must it be better to be 1 out of 6 than 2 out of 100 ... or 1,000, or 1,000,000!!!

Computations take into account the size of the Geo...

= Relative instead of Absolute Uniqueness

Figure 26

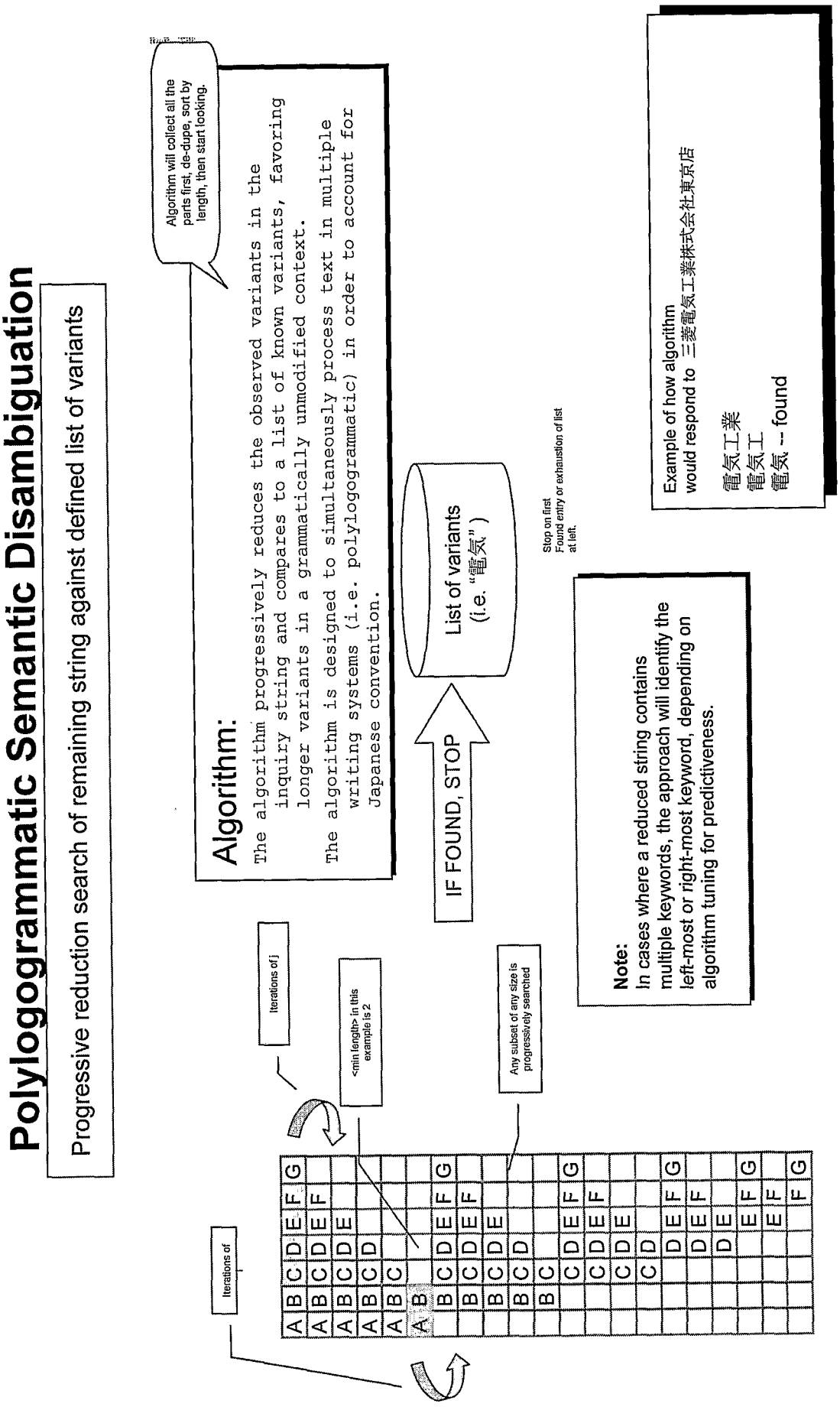

Polylogogrammatic Semantic Disambiguation
Progressive reduction search of remaining string against defined list of variants
Algorithm:
The algorithm progressively reduces the observed variants in the inquiry string and compares to a list of known variants, favoring longer variants in a grammatically unmodified context.
The algorithm is designed to simultaneously process text in multiple writing systems (i.e. polylogogrammatic) in order to account for Japanese convention.
Algorithm will collect all the parts first, de-dupe, sort by length, then start looking.
IF FOUND, STOP
List of variants (i.e. "電気")
Stop on first Found entry or exhaustion of list at left.
Iterations of
Iterations of j
<min length> in this example is 2
Any subset of any size is progressively searched
Note:
In cases where a reduced string contains multiple keywords, the approach will identify the left-most or right-most keyword, depending on algorithm tuning for predictiveness.
Example of how algorithm would respond to 三菱電気工業株式会社東京店
電気工業
電気工
電気 -- found

Figure 27

REPRESENTATIVE INDUSTRY SCORING

Scenario	MG Code
Right-most or left-most, Longest (ie most specific) Keyword generated matches reference candidate's keyword (see Figure 26 for rationale)	A
Other (shorter) keyword matches reference candidate's keyword	B
No keyword generated for inquiry matches candidate's keyword	F
No keyword generated in either inquiry or candidate.	Z

Ref	三井電気通信工業 MITSUI DENKI TSUUSHIN KOUGYOU Mitsui Electric Communications Industry	**Keyword**	電気, 通信 ⇒ 通信 Communications
Inquiry	三井情報通信 MITSUI JOUHOU TSUUSHIN Mitsui Information Communications	**Keywords**	情報通信 通信 = A 情報