代替的言語判定手法を活用した言語特定クロウラーの効率化

新井 裕樹　中平 勝子　Pann Yu Mon　三上 喜貴
長岡技術科学大学

1．はじめに

Web 上から言語特定で書かれた情報のみを収集する場合，言語判定を高速，高精度で行うことのできる言語判定技術が必須である．通常，それは HTML 本文のテキストの分析によって行われるが，一般に判定処理のために時間を要する．このため，筆者らは，より高速な代替的補助的言語判定手段の幾つかについて，その判定可能性，精度などを評価してきた．

例えば，META タグ情報である charset 情報も，言語特定だけに使用される文字であって文字コード標準が確立されている場合には有力な手掛りとなりうる[1]．田村らはその有効性について検討している[2]．一方文字コード標準が未確立の場合，現地ベンダー等は独自の非標準文字コードとこれに対応した専用フォントを開発することになり，南アジアや東南アジアの文字については特にこの傾向が見られる[3]．この場合には charset 名は確立された名前がないので記述のゆれが大きくなり，むしろフォント名が判定の決め手となる．実際この手法を活用したコーパス収集の事例も報告されている[4][5]．こうしたことから，本稿では，*n*-gram による言語判定とフォントによる言語判定を併用した言語特定クロウラーの効率化について考察する．

2．検討方法

本稿で述べる研究において，言語特定クロウラーの効率化は，ページの言語判定を構成する処理部分を高速化させ，ページの収集速度を向上させることで達成する．従来の言語特定クロウラーにおける言語判定方法としては，テキストの *n*-gram 統計量を用いる方法が一般的である．この方法は言語判定に比較的高い精度を期待できるものの，一般的に言語特定処理のために時間を要する．

筆者らが検討する代替的言語判定手法による言語特定クロウラーは，テキストの *n*-gram 統計量を用いた言語判定方法 LI と，高速な言語判定が可能なフォント名を用いた言語判定方法 FLI とを併用する．具体的には，フォント名を用いた方法で判定できない場合のみ，テキストの *n*-gram 統計量を計算して言語判定を行うことで高速化を図る．フォントによる言語判定の方法は，ある特定言語にのみ使用されるフォント名リスト L_{font} を用意し，メタタグに定義されているフォント名と L_{font} のマッチングを取ることで特定言語のみに対して言語判定を行う．

本手法を導入することで，特定言語に対しては何らかの言語判定時間 t_{lang} の効率化がなされる筈である．そこで，t_{lang} の効率化を評価する．1ページの判定に要する t_{lang} (FLI+LI)を t_{FL}，1ページの判定に要する FLI のみの t_{lang} を t_f，1ページの判定に要する LI のみの t_{lang} を t_l，FLI による判定が可能なページの使用率を x とすると，フォントと LI を両用した場合の，t_{FL} は

$$t_{FL}=x\cdot t_f+(1-x)\cdot(t_f+t_l) \quad \cdots (1)$$

とあらわされる．LI での t_{lang} を t_l と表すと，$t_{FL}<t_l$ であれば t_{lang} の効率化は可能である．

本方式の効率化を検討するには，FLI+LI 方式と LI のみの t_{lang} をそれぞれ求め，$t_{FL}<t_l$ の関係を調査し，理論値，実測値を比較する．そのためには，1）Web 空間上の言語特定フォント使用率，2）両手法の言語判定時間の関係，の2要素が重要となる．

1）Web 空間上の言語特定フォント使用率

フォントのうち特定の言語でしか使用されていないものを「言語特定フォント」と呼ぶ．実際に Web 上で言語特定フォントに遭遇する確率，つまり，使用率を ccTLD 別に調査した．言語特定フォントが多く存在するほど効率化の効果は大きい．使用率の調査に用いたデータは，Ubicrawler[6]でアジア地域（日本，中国，韓国を除く）とアフリカ地域の ccTLD より 2006 年 6 月～7 月に取得した約 1 億ページの中からサンプル抽出した 450 万ページのデータである．対象となる ccTLD は，.th(タイ語)，.vn(ベトナム語)，.lk(シンハラ語)，.in(ヒンディー語)とした．

2）両手法の言語判定時間の関係

（1）式より FLI の t_{lang} が LIM の t_{lang} に比べて短いほど効率化の寄与は多くなる．調査は，FLI

Efficient language focused crawler using alternative language identification methods
Yuki Arai, Katsuko T. Nakahira, Pann Yu Mon, Yoshiki Mikami
Nagaoka University of Technology

と LIM に 1000 件のページデータを与えて，得られた判定所要時間の平均値を明らかにする．

3．結果

3．1　理論値における評価

次に，FLI+LI，および LI のみの t_{lang} の効率化を評価する．本稿における実験では，LI 用アプリケーションとして，言語天文台プロジェクトで開発した言語判定エンジン LIM(Language Identification Module)[7]を使用した．LIM はテキストの *n*-gram 分布に基づき使用言語，使用文字，使用文字コードの判定を行う．フォントの言語判定については筆者らが開発したモジュール FLI(Font Language Identification)を使用した．

表１は，１）を評価するための ccTLD 別の言語特定フォントの使用率 x を求めたものである．

また，同データより t_f は 0.00433 秒，t_l は 5.37923 秒であった．これらの結果より，例えばランダムに Web サイト 10^6 ページ分の言語判定を行う場合，t_{lang} の効率化を求めると，.th では 147 時間，.in では 9 時間の言語判定時間を短縮することが理論上可能である．

3．2　実測値における評価

サンプルデータを対象に FLI+LI，および LI を実際に行った結果から効率化の寄与を評価する．サンプルデータは，2 つのドメイン（.th と.in）を対象に 1000 ページずつ実測した．表２に LI のみ，FLI+LI の言語判定に要した時間をまとめる．t_{FL} の実測値をA，理論値をDとし，t_l の実測値をB，理論値をEとし，$t_l - t_{FL}$ の実測値をC，理論値F，最後に，理論値-実測値をGとする．この表２より，理論値程度の判定速度をあることを確認した．

言語判定精度は，FLI によって.th ではタイ語を，.in ではヒンディー語を抽出しており，理論値 x からそれぞれ 1000 ページ中，58 ページ(.th)，4 ページ(.in)が抽出できれば良く，それに対して，.th では 32 ページ，.in では 3 ページであった．また，それらのページは，すべてタイ語，ヒンディー語と正しく判定された．

表１：言語特定フォントの使用率

	.th	.vn	.lk	.in
総ページ数	991, 559	376, 084	8, 336	398, 808
言語特定フォントを持つページ数	57, 465	5, 453	86	1, 588
x	5. 795%	1. 450%	1. 032%	0. 398%

表２：実測値，理論値の結果

単位：[秒]	.th	.in
A. t_{FL} (実測値)	8, 894	9, 372
B. t_l (実測値)	9, 146	9, 388
C. Δ_A (B - A)	252	16
D. t_{FL} (理論値)	8, 734	9, 234
E. t_l (理論値)	9, 267	9, 267
F. Δ_B (E - D)	533	33
G. Δ_C (F - C)	281	17

4．まとめ

本稿では，*n*-gram による言語判定とフォントによる言語判定を併用した言語特定クロウラーの効率化の効果を示した．これらの結果から言語特定フォントによる判定が可能なページの使用率は少ない，FLI の言語判定所要時間は LI による言語判定方法に比べると高速であることが分かった．また，理論値，実測値からフォントを用いることで言語特定クロウラーの効率化が図れることが分かった．

しかし，フォントで言語判定する際に使用する言語特定フォントリストの不備や，使用率の調査に用いたデータが Web 空間の特定の領域(ccTLD のみ)しか対象としていない点から，最大限の効果を得ることが可能な条件でないと考えられる．従って，今後これらの不足点を改善し，追加実験を行う．

参考文献

[1] IANA の charset 登録リストを見ると，言語の特定が可能な文字コード標準として日中韓やタイ語等がある．

[2] 田村孝之他，言語特定で記述された Web ページの選択的収集手法とその評価，電気情報通信学会論文誌 Vol.89-D No.2 pp.199-209

[3] 三上喜貴，文字符号の歴史：アジア編，p.205，2002

[4] Prasad Pingali, et. al., WebKhoj: Indian language IR from Multiliple Character Encodings, In Proceedings of the 15th International World Wide Web Conference, Edinburgh, Scotland, pp.801-809, 2007.

[5] Asif Ekval, Development of Bengali Named Entity Tagged Corpus and its Use in NER System, In Proceedings of the 6th Workshop on Asian Language Resources, Hyderabad, India, pp.1-8, 2008.

[6] Ubicrawler http://law.dsi.unimi.it/

[7] I. Suzuki, et.al: A Language and Character Set Determination Method Based on N-gram Statistics, ACM Transactions on Asian Language Information Processing, 1(3), pp.270-279, 2002.